AD_____

Award Number: DAMD17-98-1-8256

TITLE: Individualized Strategies for Breast Cancer Surveillance Based on Aggregated Familial Information

PRINCIPAL INVESTIGATOR: A. Yakovlev, Ph.D.
K. Boucher, Ph.D.
A. Tsodikov, Ph.D.
R. Kerber, Ph.D.
G. Gregori, Ph.D.

CONTRACTING ORGANIZATION: University of Utah
Salt Lake City, Utah 84102

REPORT DATE: July 2002

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2002	Final (1 Jan 99 - 30 Jun 02)

4. TITLE AND SUBTITLE
Individualized Strategies for Breast Cancer Surveillance Based on Aggregated Familial Information

5. FUNDING NUMBERS
DAMD17-98-1-8256

6. AUTHOR(S): A. Yakovlev, Ph.D.
K. Boucher, Ph.D.
A. Tsodikov, Ph.D.
R. Kerber, Ph.D.
G. Gregori, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Utah
Salt Lake City, Utah 84102

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Report contains color

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

This final report is concerned with stochastic modeling of breast cancer detection and estimation problems associated with the two-variate distribution of age and tumor size at diagnosis. Our methodological approach is designed to accommodate generally-structured data. Another research avenue was related to optimal scheduling of breast cancer screening by maximizing the expected reduction of tumor size at detection. We developed a Monte-Carlo EM algorithm for estimation of biologically meaningful parameters incorporated into the joint distribution of age and tumor size at detection. The proposed estimation techniques were tested by computer simulations and applied to epidemiological data on individuals identified through the Utah Population Database (UPDB) and Utah Cancer Registry. We studied various indicators of family history and used one of them to stratify the data on breast cancer obtained from the UPDB. An optimal schedule has been constructed for low- and high-risk groups of individuals identified through the UPDB. While the efficacy of the optimal schedule tends to be higher in high-risk families, its structure appears to be robust to variations in breast cancer risk.

14. SUBJECT TERMS	15. NUMBER OF PAGES
breast cancer, optimal surveillance, individualized strategies	94
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

This project is concerned with theoretical methods for designing individualized optimal strategies of breast cancer surveillance. The problem of optimal cancer surveillance is set up as a search for optimal scheduling of screening examinations subject to certain constraints on the number and timing of medical tests. The hypothesis to be tested is that the efficacy of breast cancer detection can be enhanced through incorporating aggregated family history information into a mathematical model designed to construct optimal schedules of cancer surveillance. The proposed methods have been validated using epidemiological data on breast cancer from the Utah Population Data Base (UPDB) linked to the Utah Cancer Registry (UCR). All tasks included in the Statement of Work have been addressed.

2. Modeling cancer detection

Let T be the age at tumor onset, and W the time of spontaneous tumor detection measured from the onset of disease. Introduce the random variable (r.v.) V to represent tumor size at spontaneous detection. Then $V = f(W)$, where $f : [0, \infty) \to [1, \infty)$ is a deterministic function describing the law of tumor growth. It is assumed that

(1) random variables T and W are absolutely continuous and independent;
(2) function f is differentiable and $f' > 0$, with the inverse of f denoted by g;
(3) the rate of spontaneous tumor detection is proportional to the current tumor size with coefficient $\alpha > 0$.

The probability density function (p.d.f.) of the vector $Y = (T + W, V)$ is given by

$$p_Y(u, v) = p_T(u - g(v))p_V(v). \tag{1}$$

In the particular case of exponential tumor growth with rate $\lambda > 0$ we have

$$p_Y(u, v) = \frac{\alpha}{\lambda} e^{-\frac{\alpha}{\lambda}(v-1)} p_T(u - \frac{\ln v}{\lambda}), \quad u \geq 0, \ 1 \leq v \leq e^{\lambda u}. \tag{2}$$

It has been proven (Hanin, 2001) that the distribution (2) is identifiable if the density p_T for the time of tumor onset is specified by the Moolgavkar–Venzon–Knudson (MVK) two-stage model of carcinogenesis. However, the shape of the marginal distribution of tumor volume obtained from (2) is inconsistent with actual observations. Therefore, a reasonable generalization of the above distribution is necessary.

To make the distribution (2) more flexible, suppose that the process of tumor growth is described by the exponential law $f(w) = e^{\lambda w}$, $w \geq 0$, with a *random* growth rate λ. Specifically, we assume that the parameter $1/\lambda$ is gamma distributed with parameters a and b. Then we have

$$p(u, v) = \frac{\alpha b^a}{\Gamma(a)} \int_0^{u/\ln v} t^a \exp\{-[b + \alpha(v - 1)]t\} p_T(u - t \ln v) dt$$

$$= \frac{\alpha b^a}{(\ln v)^{a+1}\Gamma(a)} \int_0^u (u - s)^a \exp\{-\frac{b + \alpha(v - 1)}{\ln v}(u - s)\} p_T(s) ds, \tag{3}$$

for $u \geq 0$, $v \geq 1$. Unfortunately, no identifiability results on this distribution are available because the corresponding theoretical problem is far too difficult to approach. Once the density p_T of the age at tumor onset T is specified within a certain parametric family, equation (3) allows us to compute p.d.f. of the joint distribution of age and tumor size at detection. The above model and its properties are discussed at length in our recent paper (Bartoszyński et al., 2001) included in Appendix 1.

3. The estimation procedure.

For practical purposes, the joint distribution can be reparametrized to include the tumor diameter rather than its volume. If $U = T + W$ is the time (age) at tumor detection it is easy to see that the joint p.d.f. of (U, D) is given by

$$p(u,d) = \frac{3\alpha d^2 b^a}{d_0^3 \Gamma(a)} \int_0^\infty \theta^a e^{-\theta[b+\alpha((\frac{d}{d_0})^3-1)]} p_M(u - 3\theta \ln(d/d_0))\, d\theta, \tag{4}$$

where d_0 is the diameter of a single tumor cell ($d_0 \approx 10^{-3}$ mm) and p_M is the p.d.f. of the MVK distribution. The model depends on six parameters: the rate of detection α, the mean $\mu = a/b$ and standard deviation $\sigma = \sqrt{a}/b$ of the gamma distribution, and the identifiable parameters A, B, ρ of the MVK distribution. The following explicit formula holds for the marginal distribution of the diameter:

$$p(d) = \frac{3d^2 \alpha a b^a}{d_0^3 [b + \alpha((d/d_0)^3 - 1)]^{a+1}}. \tag{5}$$

It is important to note that, while three parameter appear in (5) they are not all identifiable from data on tumor diameters. In particular, we can rewrite (5) in terms of the two parameters $\beta = \alpha\mu$ and $\delta = \mu/\sigma$ as

$$p(d) = \frac{3d^2 \beta}{d_0^3 [1 + \frac{\beta}{\delta^2}((\frac{d}{d_0})^3 - 1)]^{\delta^2+1}}. \tag{6}$$

A direct maximization of the likelihood generated by formula (4) in the presence of censoring, data truncation, and/or missing size information, is practically infeasible. Each of the situations mentioned above lead to the presence in the likelihood of complicated double integrals (many thousands of them for the datasets of interest) and their computation is difficult to manage from the point of view of both computation time and error control. We overcome the obstacle by using a Monte Carlo EM (MCEM) algorithm, first proposed by Wei and Tanner (1990) (see also McLachlan and Krishnan,1997; Chan and Ledolter, 1995).

In the standard EM algorithm, the E step consists in computing the conditional expectation of the complete data log–likelihood given the observed data. In the MCEM algorithm the conditional expectation of the log–likelihood of the complete data is estimated by averaging the conditional log-likelihoods of simulated sets of complete data. The MCEM algorithm does not possess the same monotone convergence properties as the standard EM algorithm, however it is shown in (Chan and

Ledolter, 1995) that, under suitable regularity conditions, an MCEM sequence will, with high probability, get close to a maximizer of the likelihood of the observed data.

Let $Y = y$ be the observed incomplete data, $Z = z$ the missing data, and $X = x$ the unobserved complete data, with $x = (y, z)$. Let ψ be an arbitrary element in the parameter space, $E_\theta(\cdot|y)$ denotes the conditional expectation given $Y = y$ with θ treated as parameter, and $l_X(\theta')$ is the log-likelihood of X. Given a sample $z_1(\theta), \ldots, z_m(\theta)$ from $p_\theta(z|y)$, the conditional pdf of Z given $Y = y$ and θ, the MCEM algorithm consists of two steps:

$$\text{MCEstep}: \quad Q(\theta'|\theta) = \frac{1}{m}\sum_{i=1}^{m} l_{(z_i(\theta),y)}(\theta'),$$

$$\text{Mstep}: \quad \text{maximize } Q(\cdot|\theta).$$

A stopping rule for the MCEM algorithm and a discussion of the effect of the size m of the sample $z_i(\theta)$ can be found in (Chan and Ledolter, 1995).

In our particular situation, the complete data can be represented as $X(\xi) = (Y(\xi, \delta(\xi)), Z(\xi, \delta(\xi)))$, where $\delta(\xi)$ is a discrete random variable that takes the values 0 when the observation is censored, 1 in case of a failure for which we measure the size of the tumor, and 2 for a failure when the tumor size is not recorded. If we let T_c be the time of censoring, we can then take

$$Y(\xi, 0) = (T_c(\xi)), \quad Z(\xi, 0) = (T(\xi), \theta(\xi)), \quad \text{when } \delta(\xi) = 0,$$

$$Y(\xi, 1) = (U(\xi), D(\xi)), \quad Z(\xi, 1) = 0, \quad \text{when } \delta(\xi) = 1,$$

and

$$Y(\xi, 2) = (U(\xi)), \quad Z(\xi, 2) = (T(\xi), \theta(\xi)), \quad \text{when } \delta(\xi) = 2.$$

Of course, in the last case knowing U,T, and θ, implies knowing $W = U - T$ and D from the law of tumor growth (which is exponential in our case). The log–likelihood for X can easily be derived.

The choice of initial values for the six parameters of the model is reduced to a one–dimensional problem by providing a preliminary fit of the tumor size data by the marginal distribution (6) and separately of the age data by the MVK distribution. This allows us to obtain a starting point estimate for five out of the six parameters incorporated into the model. In particular we can proceed by choosing the rate of detection α as the only parameter which needs to be assigned an arbitrary initial value.

4. The data

The study population consisted of people recorded in the Utah Population Database, who were born between 1936 and 1941 and for whom follow-up information is available that places them in Utah during the years of operation of the UCR. The analysis was performed on subcohorts based on birth year. In particular we looked at five separate cohorts for female breast cancer. As the UCR has records post 1965 only,

Table 1: Data description.

Birth Cohort	Sample Size	# Failures	# Missing Size Obs.
1918-23	16672	960	368
1924-29	15032	804	300
1930-35	12882	576	185
1936-41	11374	410	96
1942-47	13437	333	79

there is a left truncation effect, different for different birth cohorts, which is taken in account by our algorithm.

The UCR has been recording cancer size since 1975. The latest data (post 1987) offer the tumor diameter at the time of diagnosis in millimiters, while older records only give coarser intervals. In order to simplify the analysis we have decided to remove the grouping of the data by treating them as uniformly distributed in each interval.

The relevant information for each birth cohort in our population is given in Table 1. Here we denote by failures the cases where we have tumor size information and report as missing data the few cases where the presence of breast cancer was known but no size data was available.

5. Data analysis

To adjust for birth cohort effects we extended the model (4) to allow the parameter ρ in the MVK model be a function of the birth cohort. This is equivalent to using the proportional hazards model with the birth cohort incorporated as a categorical covariate. We first applied our estimation procedure to one-year subcohorts (B36, B37, ..., B41) comprising the cohort of individuals born between 1936 and 1941 (Table 2).

Table 3 presents the maximum likelihood estimates of model parameters that result from our estimation procedure when fitting the model to the birth subcohort data described in Table 2. The resultant fit to the marginal distributions of U and D is shown in Figures 1–4. From Table 3, it is clear that the parameter ρ does not very significantly among the sub-cohorts under study. This observation allowed us to group birth cohort data in 5 year intervals in further data analyses required to finalize the project (see below). It should be noted that the likelihood profile with respect

Table 2: Data description.

Birth Cohort	Sample Size	# Failures	# Missing Size Obs.
B36	1911	87	29
B37	1932	73	16
B38	1902	76	14
B39	1870	50	11
B40	1896	63	20
B41	1925	66	16

to α is very flat, thereby deteriorating the accuracy of estimation of the sensitivity parameter α. Our computer simulations, conducted at different (fixed) values of the parameter α, have demonstrated that the proposed procedure produces good estimates of the product $\alpha\mu$ and the ratio μ/σ; even in moderate sample studies these estimates are stable numerically and appear to be fairly close to the true parameter values in a wide range of α.

The same analysis was performed on five separate birth cohorts, each encompassing a contiguous six-year period (Table 1). The resultant fit to the marginal distributions of tumor diameter and age at diagnosis is shown in Figures 5 and 6; the corresponding parameter estimates are given in Table 4. The model provided an excellent description of all the cohorts under study. The mutual dependence of U and D, as captured by the expected tumor size conditional upon age at detection, is also consistent with the data.

6. Optimal screening schedules

The sequence of moments of time assigned for medical exams and counted from the birth of a patient are called a *screening schedule*. Let $\mathcal{T}$ be the set of all possible screening schedules $\tau = \{\tau_1 < \tau_2 < ... < \tau_n\}$. The set $\mathcal{T}$ may be subject to (some of) the following restrictions:

(a) $n \leq n_0$, where n_0 is an upper bound for the number of exams;

(b) $\tau_1 \geq m$ and $\tau_n \leq M$, where m and M are the earliest and the latest times for the first and the last exams, respectively;

(c) $\tau_{i+1} - \tau_i \geq h > 0$ for all $i = 1, 2, ..., n-1$. This condition suggests a lower bound h for the minimal duration between any two successive exams.

Table 3: Maximum likelihood estimates of model parameters for birth sub-cohorts B36-B41

Parameter	Estimated Value
α	$2.00\ 10^{-13}$
μ	1.01
σ	$1.19\ 10^{-1}$
A	$2.51\ 10^{-1}$
B	$6.63\ 10^{-6}$
$\rho(36)$	$1.47\ 10^{-2}$
$\rho(37)$	$1.17\ 10^{-2}$
$\rho(38)$	$1.28\ 10^{-2}$
$\rho(39)$	$9.19\ 10^{-3}$
$\rho(40)$	$1.35\ 10^{-2}$
$\rho(41)$	$1.39\ 10^{-2}$

Table 4: Maximum likelihood estimates of model parameters for five birth cohorts

Parameter	Estimated Value
α	$2.00\ 10^{-13}$
μ	1.01
σ	$9.09\ 10^{-2}$
A	$1.45\ 10^{-1}$
B	$6.04\ 10^{-5}$
ρ(18-23)	$2.94\ 10^{-2}$
ρ(24-29)	$3.65\ 10^{-2}$
ρ(30-35)	$3.33\ 10^{-2}$
ρ(36-41)	$4.21\ 10^{-2}$
ρ(42-47)	$5.11\ 10^{-2}$

Fig. 1

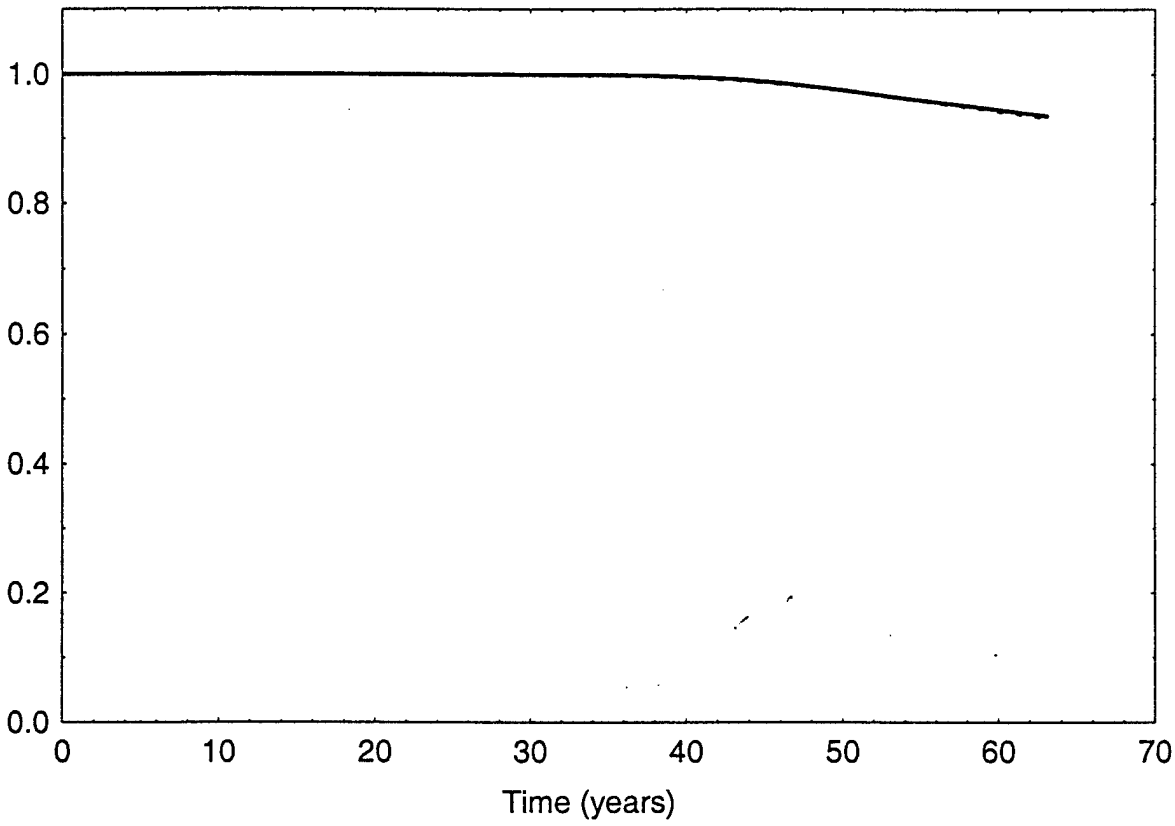


Figure 1. Parametric versus non–parametric estimates of the survivor function for one sample sub-cohort(B36). The solid line represents the Kaplan–Meier estimate, while the dotted line gives the model-based parametric estimate. The two curves are practically identical.

Fig. 2

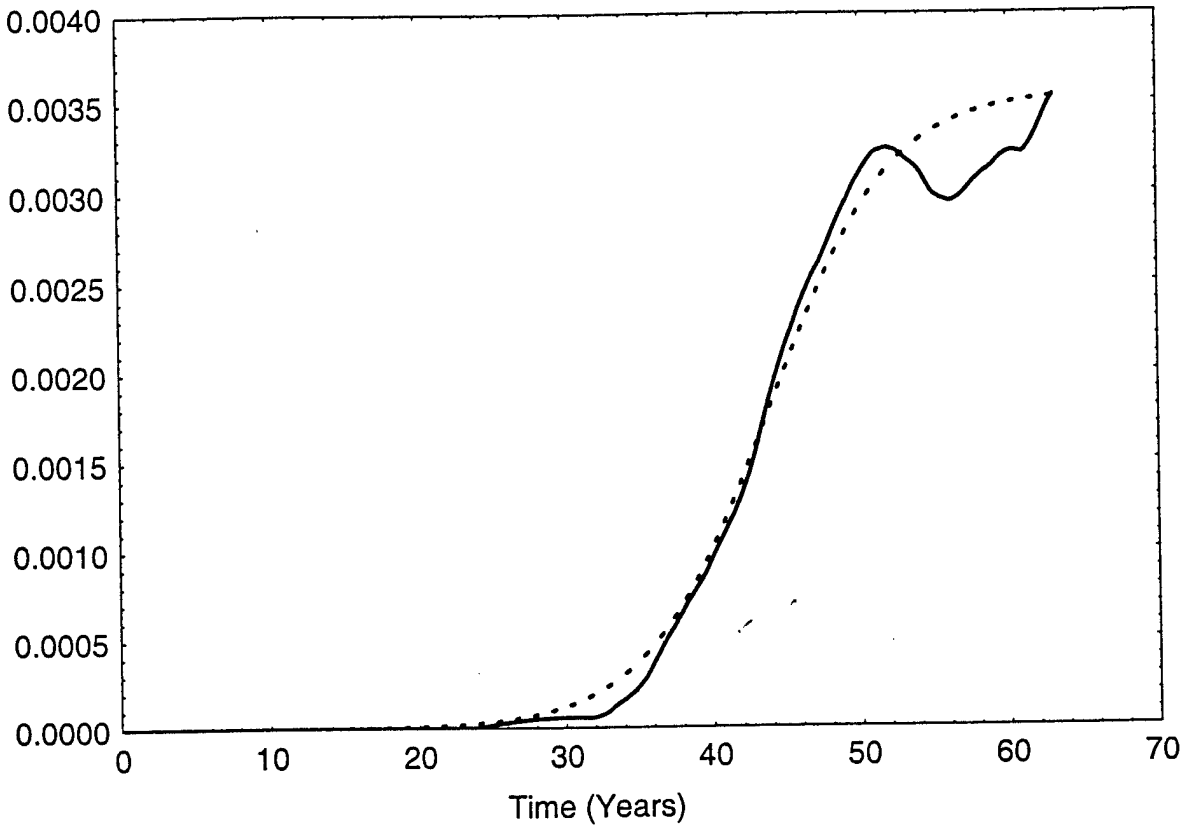


Figure 2. Parametric versus non–parametric estimates of the hazard function for one sample sub-cohort(B36). The solid line represents a local likelihood kernel-smoothed estimate, while the dotted line shows the model-based parametric estimate.

Fig. 3

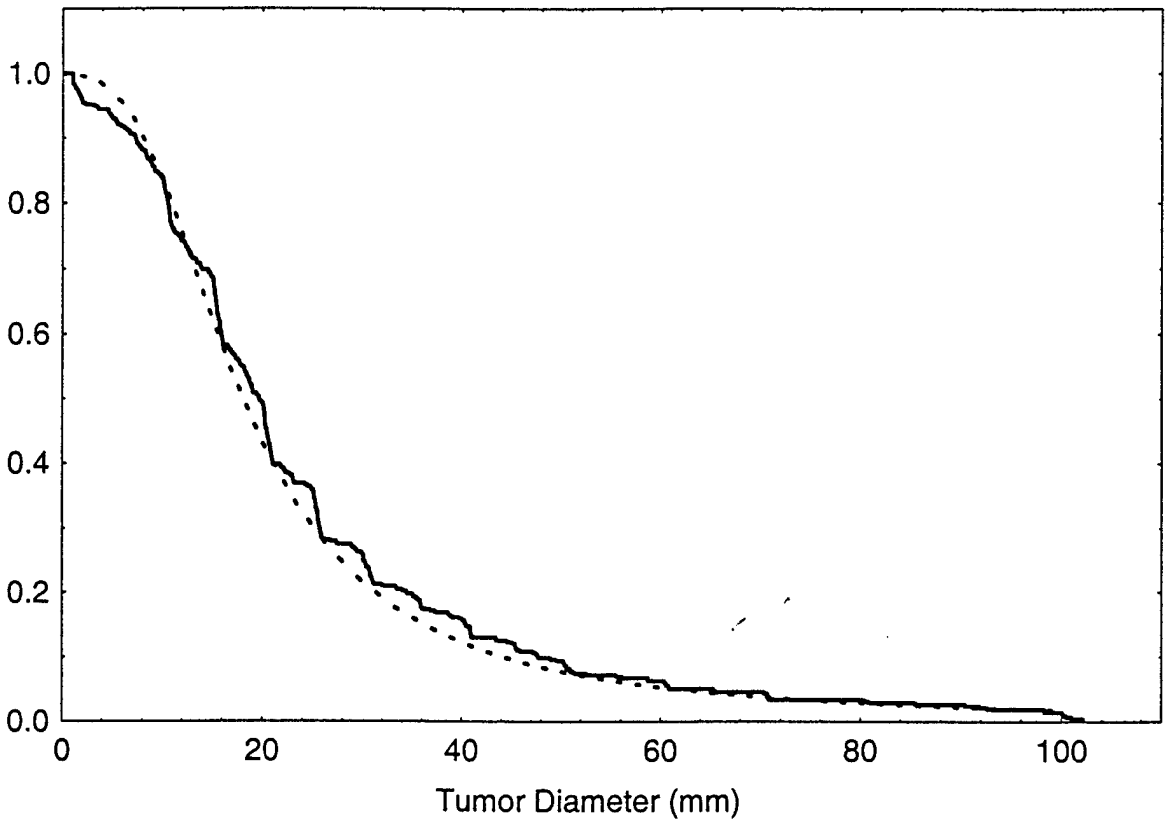


Figure 3. Parametric versus non–parametric estimates of the tail function of the tumor diameter at diagnosis for the total population studied (birth years 1936 through 1941). The solid line represents the Kaplan–Meier estimate, while the dotted line gives the values computed using our model.

Fig. 4

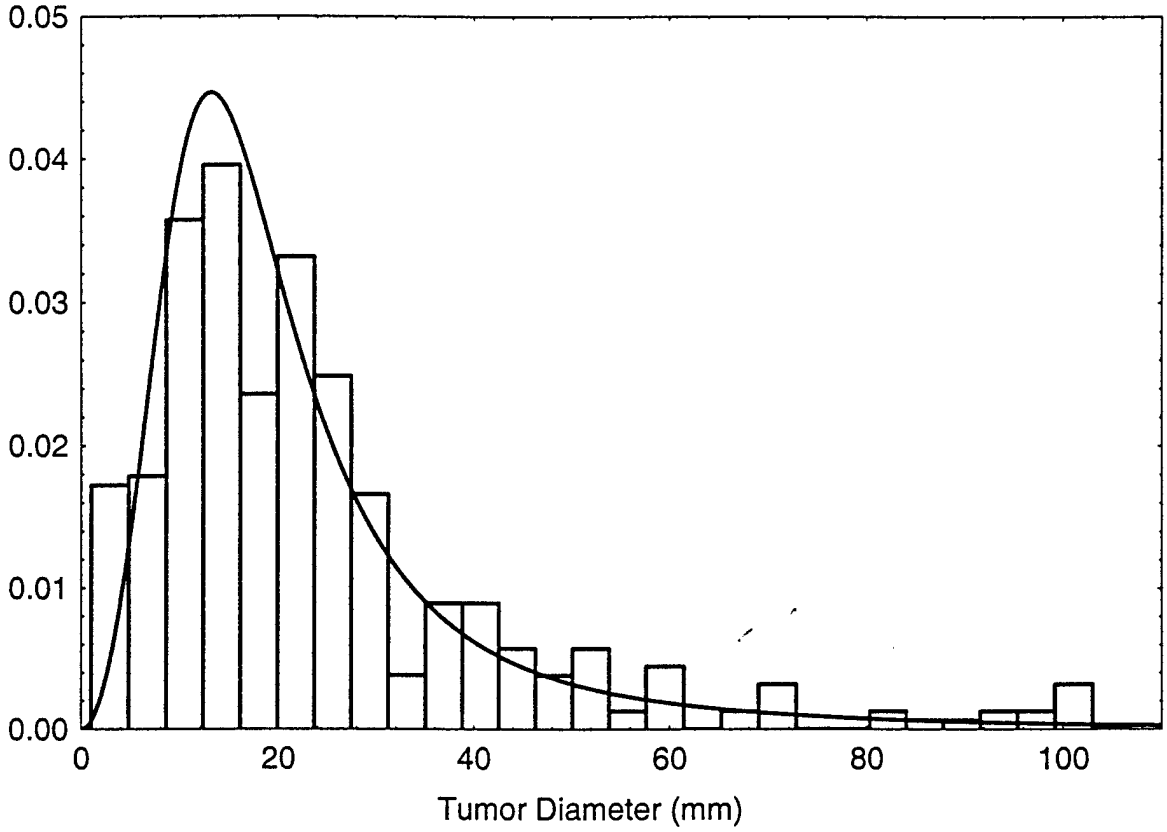


Figure 4. This figure shows a histogram for the tumor diameter at diagnosis of the total population studied (birth years 1936 through 1941) and the corresponding probability density predicted by the model.

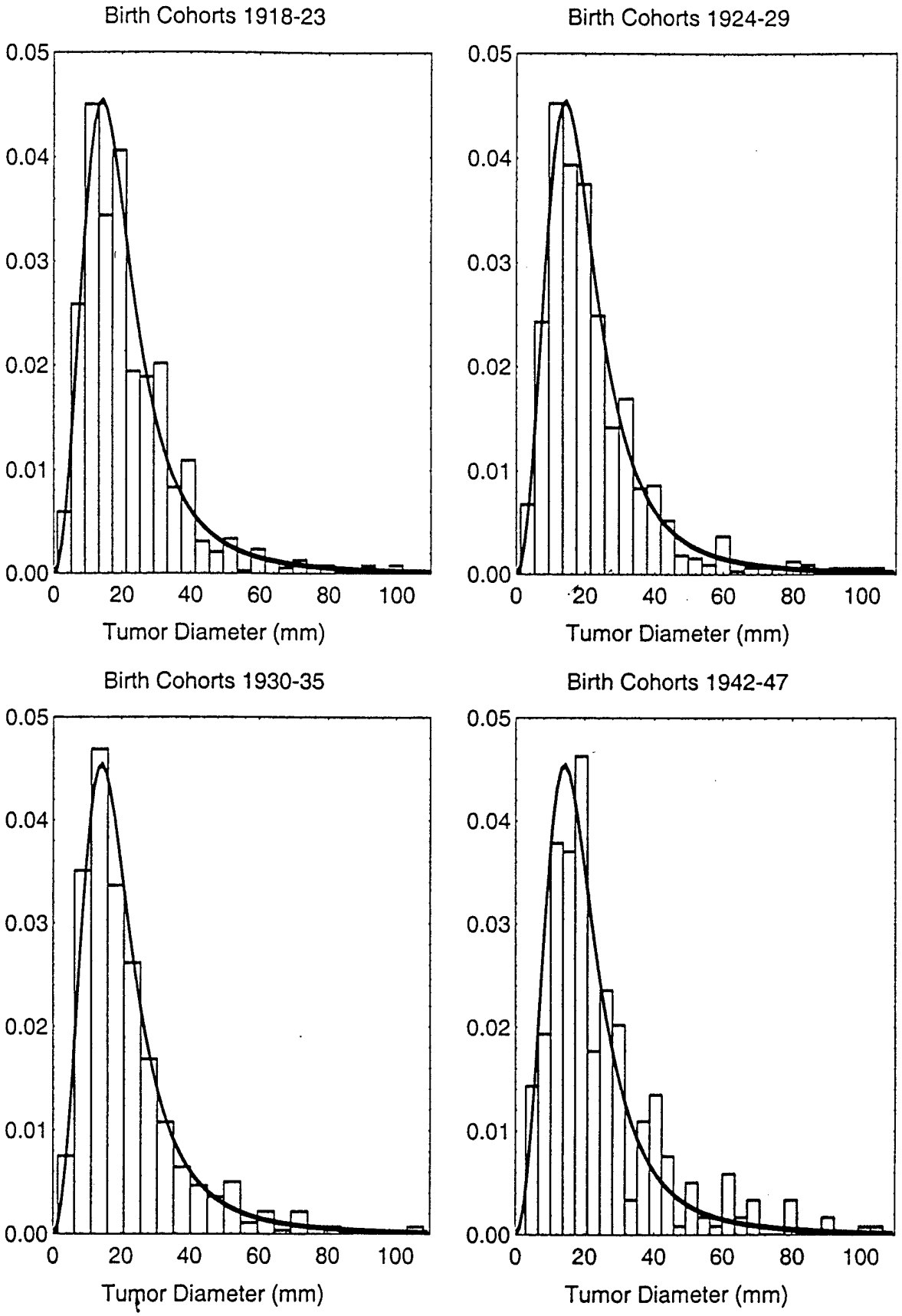


Figure 5. Distribution densities for tumor size at detection.

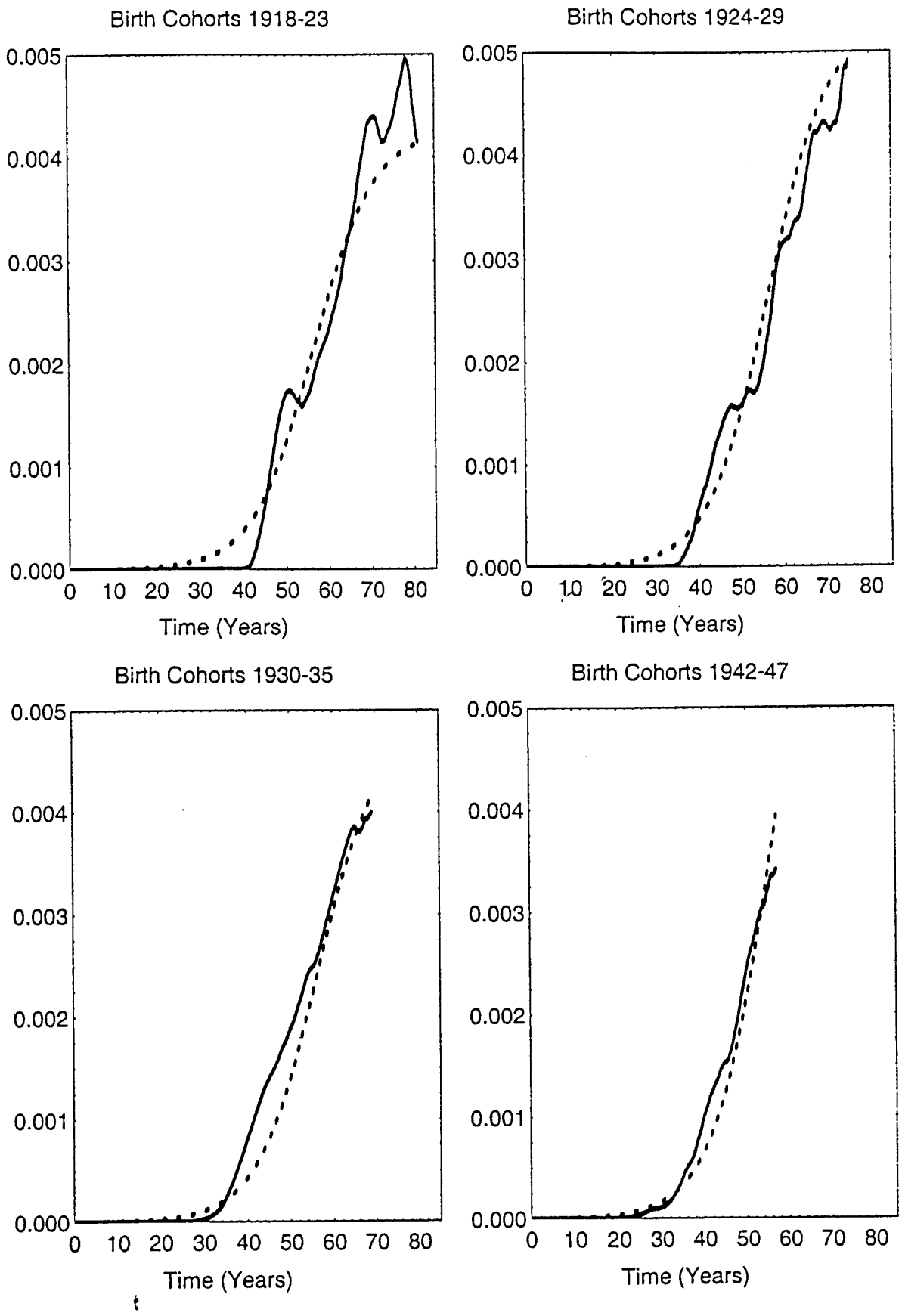


Figure 6. Hazard functions for age at detection. Solid lines: local likelihood kernel-smoothed estimates, dotted lines: model-based parametric estimates.

We distinguish between *spontaneous* (incidental) and *screening based* tumor detections. The first occurs in the absence of or concurrently with screening and is thought of as a continuous process. In contrast to this, screening based detection is an instantaneous event that may occur only at the moments of the prescribed medical exams and is therefore a discrete process. When both types of detection are present, they can be viewed as competing risks.

For a discrete screening schedule $\tau \in \mathcal{T}$, we define the efficiency functional as the Kantorovich distance between the tumor sizes S_0 and S at spontaneous and combined detection:

$$d(S, S_0; \tau) = \int_1^\infty \mid \bar{G}_{S_0}(s) - \bar{G}_S(s) \mid ds.$$

Since $\bar{G}_{S_0}(s) \geq \bar{G}_S(s)$ for all $s > 1$,

$$d(S, S_0; \tau) = E\{S_0\} - E\{S\},$$

where $E\{.\}$ stands for the expectation.

Suppose the law of tumor growth is given by $f_\theta(t)$ and θ is non-random. Then

$$d(S, S_0; \tau) = \sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} \sum_{j=i+1}^{n} e^{-\alpha \sum_{k=i+1}^{j-1} f_\theta(\tau_k - t)}$$
$$\times [1 - e^{-\alpha f_\theta(\tau_j - t)}] R(\tau_j - t) dG_T(t),$$

where

$$R_\theta(x) := \int_x^\infty \bar{G}_{W_0}(w) f'_\theta(w) dw, \quad x \geq 0.$$

The extension of this formula to the case of random θ is straightforward (see the paper by Hanin et al., 2001, included in Appendix 4).

7. The Effects of Family History

7.1. Estimation of the hazard rate

Proceeding from preliminary studies of different spline estimation procedures, we chose to model the hazard function via quadratic splines. A quadratic spline with m knots specifies the hazard to be of the form

$$\lambda_m(t) = \sum_{i=0}^{2} \gamma_{0i} t^k + \sum_{j=1}^{m} \gamma_{j2} (t - \tau_j)_+^2 \tag{1}$$

where $(x)_+ = \max(x, 0)$. For each birth cohort, we fit splines with knots which are equally spaced in the interior of the interval $[T_{min}, T_{max}]$, where T_{min} is the minimum truncation age in the cohort and T_{max} the maximum follow-up (failure or censoring)

time. Restrictions are placed on the coefficients to ensure that $\lambda_m(t)$ remains positive for all t. Thus with m knots the number of parameters is $m + 3$. Models can be fit using maximum likelihood techniques applied to the corresponding conditional likelihood, as discussed in Boucher and Kerber (2001a).

We have developed software designed to compute the spline estimates by maximizing the likelihood function using the algorithm of Powell. We start with one knot and increase the number of knots until the fit is not improved, as determined by the likelihood ratio test at the significance level $\alpha = 0.05$. Three other subcohort estimates of the hazard function were computed for comparison with the spline estimator; an estimator of the life table type, a Gaussian kernel estimate based on the Nelson-Aalen nonparametric estimator, and local likelihood estimators with different kernels (uniform, Epachechnikov, and Gaussian). All the estimators mentioned above are in good agreement with each other when applied to the UPDB data.

Using the computer programs developed in Year 1, the hazard function for cancer incidence has been estimated from left truncated and right censored data on individuals identified through the UPDB and UCR.

Although the estimates become less reliable at increasing age, the hazard function for breast cancer appears to be essentially non-decreasing in all the categories of all familial measures considered. Thus we find no evidence of an "immune fraction" in this analysis. The curves for different levels of risk appear not to merge or cross, indicating that the increased risk to those with a family history does not dissipate after a certain age.

This study is presented at length in the paper by Boucher and Kerber (2001a) included in Appendix 2.

7.2. Measures of Familial Aggregation as Predictors of Breast Cancer Risk

Several measures of familial disease aggregation have been proposed, but only a few of these are designed to be implemented at the individual level. We have evaluated four of them in the context of breast cancer incidence. After extensive discussions, we came to the conclusion that testing different measures of family history with simulated data was not warranted in view of the fact that such a study would have added little to the results of real data analysis. Therefore, we decided to focus on a more comprehensive analysis of epidemiological data employing a wider spectrum of potential predictors of breast cancer risk.

A population-based cohort consisting of 114,429 women born between 1874 and 1931 and at risk for breast cancer after 1965 was identified by linking the UPDB and the UCR. Three competing methods were used to obtain predictors of familial aggregation of risk: the number of first degree relatives with breast cancer, the posterior probability of carrying BRCA1 or BRCA2, and the Familial Standardized Incidence Ratio (FSIR), which weights the disease status of relatives based on their degree of relatedness with the proband. Spline regression methods were used to estimate the hazard function, stratified by measures of familial aggregation.

We dichotomized each of our measures of familial risk, with the high risk category

representing approximately 8.5% of the data in each case. This was a natural cut point, as it represents the proportion of subjects with one or more first degree relatives with breast cancer. The cutoff for FSIR roughly corresponds to a relative risk of two to family members. The cut points for the posterior probability of BRCA1 and BRCA2 come at points where the posterior probability is rather small, less than 0.0005 in both cases.

Our previous analysis indicated that a highly significant birth-year effect exists in the data, with a women born ten years later having an estimated 40% increased age-specific risk. Birth-year was included as an additional covariate in all regression analyses. The baseline risk was estimated using splines, with the proportional hazards model used for birth-year and familial risk. As with most of the models, we found that two knots were sufficient to provide an optimal fit.

The presence of a first degree relative with breast cancer and the dichotomized FSIR variable each appear to be equally effective at distinguishing high risk subjects, with the high risk category having about double the risk, while the posterior probability of BRCA1 and BRCA2 appear to be less effective.

We performed a more detailed stratified analysis of FSIR. The category boundaries were the approximate 75th, 90th, and 99.9th percentiles of the (adjusted) FSIR distribution. The upper category roughly corresponds to the reported fraction of the general population carrying known breast cancer genes. Bootstrap confidence bands were computed as well as an indicator of the reliability of the estimates. The bootstrap confidence intervals are based on 100 bootstrap samples, except for the $<$ 75th percentile category, which is based on 20 bootstrap samples, because of the extensive time it took to fit the models to the large datasets.

We incorporated the posterior probabilities of BRCA1 and BRCA2 and their logarithms, as well as $\log\log FSIR$ as continuous variables in separate analyses, using a proportional hazards model with birth-year as an additional covariate. The best result (in terms of statistical significance) was obtained by including the $\log\log FSIR$, where we get a likelihood ratio $\chi^2_1 = 316.72$ ($p < 0.00001$).

We also considered the indicator variable NFIRST for presence/absence of a first degree relative, in a proportional hazards model. The behavior of the hazard function across different strata shows that the proportional hazards assumption is not grossly violated. The variable NFIRST was highly significant (likelihood ratio $\chi^2_1 = 185.6$, $p < 0.0001$). Addition of a second indicator variable for two or more first degree relatives with breast cancer did not improve the likelihood significantly. More technical details on this study are given in the paper by Boucher and Kerber (2001b) included in Appendix 3.

7.3. Individualized strategies of optimal screening

In our analysis, the function $f_\theta(t)$ was taken to be deterministic exponential with rate λ. The reciprocal of λ was assumed to be gamma-distributed with shape parameter a and scale parameter b. To evaluate the effect of family history, the data were stratified by $FSIR$ value and all parameters of the model were estimated from each stratum. To prevent the strata from being too small, we divided the data in two

Table 5: Parameter estimates obtained from the UPDB data stratified by FSIR.

Parameter	$FSIR < 1$	$FSIR > 1.8$
α	2×10^{-13}	2×10^{-13}
μ	0.76	1.08
σ	0.074	0.118
A	0.130	0.133
B	6.13×10^{-5}	8.72×10^{-5}
ρ (1918-23)	0.034	0.045
ρ (1924-29)	0.037	0.060
ρ (1930-35)	0.044	0.064
ρ (1936-41)	0.057	0.079
ρ (1942-47)	0.048	0.087

groups defined as $FSIR < 1$ and $FSIR > 1.8$. Using the estimation procedure described in Sections 4 and 5 we obtained estimates of the basic parameters for each stratum (Table 5).

In each stratum, the search for optimal screening schedules and optimal screening efficiency was conducted for a fixed number $n = 10$ of screens with no restriction on the moments of exams. The period of observation (horizon) was truncated at 100 years. The method of optimization was the exhaustive search with the step of 0.25 years.

For both groups, our algorithm results in the same optimal schedule represented by the following sequence of screening ages (years): $\tau_0 = 74.5, \tau_1 = 78, \tau_2 = 81, \tau_3 = 83.75, \tau_4 = 86.50, \tau_5 = 89.25, \tau_6 = 92, \tau_7 = 94.75, \tau_8 = 97.50, \tau_9 = 100$. For this schedule, the intervals $\Delta_i := \tau_i - \tau_{i-1}$, $i = 1, ..., n$, between two successive exams are as follows: $\Delta_1 = 3.5, \Delta_2 = 3.0, \Delta_3 = 3.0, \Delta_4 = 2.75, \Delta_5 = 2.75, \Delta_6 = 2.75, \Delta_7 = 2.75, \Delta_8 = 2.75, \Delta_9 = 2.5$. Thus the structure of the optimal schedule appears to be the same for both groups thereby indicating that individualization of screening schedules is not warranted. However, this unique optimal schedule does provide a tangible gain (relative to the absence of screening) in terms of the mean tumor volume

Table 6: The percent decrease in the mean tumor volume under the optimal screening schedule (relative to the absence of screening).

Birth Cohort	$FSIR < 1$	$FSIR > 1.8$
1918-23	0.14%	0.19%
1924-29	0.14%	0.25%
1930-35	0.16%	0.82%
1936-41	0.21%	0.32%
1942-47	0.18%	0.34%

at diagnosis, the gain being higher in the high risk group ($FSIR > 1.8$) as evidenced by the results shown in Table 6. The optimal schedule depends quite strongly on the sensitivity parameter α which needs to be more reliably estimated from similar data generated by randomized screening trials. A research paper summarizing the above results is in preparation.

8. Key Research Accomplishments

Our key accomplishments can be summarized briefly as follows:

- We have developed computer programs implementing four statistical procedures for estimation of the hazard function; these procedures accommodate data subjected to random truncation and censoring.

- A new method has been developed for designing optimal schedules of breast cancer surveillance specially adapted to population-based settings.

- Numerical experiments have shown that mathematical and computational problems of optimal cancer surveillance are tractable within the framework of the proposed model of cancer surveillance and detection.

- The joint distribution of age and tumor size at diagnosis has been derived within the framework of the proposed model of the natural history of breast cancer.

- A Monte-Carlo EM algorithm has been developed for estimation of the parameters incorporated into the joint distribution of age and tumor size at detection.

- The usefulness of the estimation procedure was evaluated by computer simulations.

- The estimation procedure was applied to epidemiological data on individuals identified through the UPDB and stratified by one of the most widely accepted indicator of family history of breast cancer (FSIR). This application provided values of model parameters to be used for evaluating potential benefits from individualized schedules.

- Given the estimated parameter values optimal schedules have been constructed using the stratified data on breast cancer. This study has shown that the optimal schedule of breast cancer screening is robust to variations in familial risk.

9. Reportable Outcomes

Publications

1. Bartoszyński, R., Edler, L., Hanin, L., Kopp-Schneider, A., Pavlova,L., Tsodikov, A., Zorin, A., and Yakovlev, A. Modeling cancer detection: Tumor size as a source of information on unobservable stages of carcinogenesis, *Mathematical Biosciences* 171: 113-142, 2001.
2. Boucher, K.M. and Kerber, R.A. The shape of the hazard function for cancer incidence, *Mathematical and Computer Modelling* 33: 1361-1376, 2001.
3. Boucher, K.M. and Kerber, R.A. Measures of Familial Aggregation as Predictors of Breast Cancer Risk, *Journal of Epidemiology and Biostatistics* 6(5): 377-385, 2001.
4. Hanin, L.G., Tsodikov, A.D., and Yakovlev, A.Y. Optimal schedules of cancer surveillance and tumor size at detection, *Mathematical and Computer Modelling* 33: 1419-1430, 2001.
5. Hanin, L.G., Identification problem for stochastic models with application to carcinogenesis, cancer detection, and radiation biology, *Discrete Dynamics in Nature and Society*, 6: 1-14, 2002.

Presentations and Meeting Abstracts

1. Yakovlev, A.Y., Tsodikov, A.D., and Hanin, L.G. Optimal schedules of breast cancer surveillance, Abstract, Era of Hope Meeting, Atlanta, June 2000.
2. Boucher, K.M. and Kerber, R.A. The shape of the hazard function for cancer incidence, Abstract, Era of Hope Meeting, Atlanta, June 2000.
3. Yakovlev, A.Y. Stochastic modeling of carcinogenesis and cancer detection, Invited presentation, Minisimposium *Cancer Modeling*, First SIAM Conference on the Life Sciences, Boston, March 7-8, 2002.

Awards

1. Grant 1 U01 CA88177-01, NIH/NCI, Mechanistic Modeling of Breast Cancer Surveillance, RFA "Cancer Intervention and Surveillance Network (CISNET)", P.I.: Yakovlev, A.Y., 09/01/00 - 08/31/04, total costs: $ 537,653.

2. Grant proposal "Quantitative insight into the natural history of breast cancer" (PI: A. Yakovlev), DOD Breast Cancer Program, 2002.

10. Conclusions

A version of the Monte Carlo EM algorithm has been developed for maximum likelihood inference based on the distribution of age and tumor size at detection. This algorithm was tested by computer simulations and an application to breast cancer data obtained from the UPDB and UCR datasets. Notwithstanding the fact that the likelihood profile with respect to the sensitivity parameter α appears to be very flat, the proposed procedure produces good estimates of the product $\alpha\mu$ and the ratio μ/σ; these estimates are quite insensitive to specific values of the parameter α.

We have explored several methods of measuring familial aggregation at the individual level as applied to breast cancer data. All prove to be significant predictors of individual risk. Judging by the difference in risk estimates, as well as the likelihood ratio test, presence of a first degree relative and FSIR appear to be better indicators of increased risk than the posterior probability of BRCA1 or BRCA2. Proceeding from these results, we used the simplest indicator, namely the presence of a first degree relative with breast cancer, for the purposes of data stratification. We constructed optimal schedules of cancer surveillance (screening) to each data stratum. The next step was comparing the optimal schedules thus obtained and evaluate their efficacy in terms of the proposed criterion of optimality.

Given the estimated parameter values an optimal schedule has been constructed using the available data on breast cancer. This schedule provides a maximum reduction of the mean tumor size at detection over a set of discrete screening schedules. While the efficacy of the optimal schedule tends to be higher in high risk families, its structure appears to be robust to variations in breast cancer risk. The optimal schedule appears to depend quite strongly on parameters characterizing the sensitivity of spontaneous and screen-based detection. More reliable estimates of these parameters are needed. This is likely to be accomplished by analyzing similar data generated by randomized screening trials. The question of whether the expected gain in the mean tumor size at diagnosis translates into a tangible survival benefit remains to be addressed in future studies.

So What?

This study shows that the efficacy of breast cancer screening (in terms of tumor size at diagnosis) varies depending on family history and genetic predisposition for breast cancer. However, this effect is not sufficiently strong to change the structure of an optimal schedule of breast cancer screening designed to maximize the reduction of tumor size at the time of detection. Further studies are needed to establish a link between the effect of screening on the distribution of tumor size at detection and post-treatment survival of patients diagnosed with breast cancer.

Personnel

1. A. Yakovlev, Ph.D. (PI)
2. K. Boucher, Ph.D. (Co-investigator)
3. R. Kerber, Ph.D. (Co-investigator)
4. A. Tsodikov, Ph.D. (Co-investigator)
5. G. Gregori, Ph.D. (Research Associate)
6. L. Hanin, Ph.D. (Consultant)

References

1. Bartoszynski R., Edler L., Hanin L., Kopp–Schneider A., Pavlova L., Tsodikov A., Zorin A., Yakovlev A., Modeling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis, *Math. Biosci.* 171 (2001) 113.

2. Boucher, K.M. and Kerber, R.A. The shape of the hazard function for cancer incidence, *Mathematical and Computer Modelling* 33: 1361-1376, 2001.

3. Boucher, K.M. and Kerber, R.A. Measures of Familial Aggregation as Predictors of Breast Cancer Risk, *Journal of Epidemiology and Biostatistics* 6(5): 377-385, 2001.

4. Chan K. S., Ledolter J., Monte Carlo EM Estimation for Time Series Models Involving Counts, *J. Am. Stat. Ass.* 90: 242-252,1995.

5. Hanin, L.G., Identification problems for stochastic models with application to carcinogenesis, cancer detection, and radiation biology, *Discrete Dynamics in Nature and Society*, 6: 1-14, 2002.

6. McLachlan G. J., Krishnan T., *The EM Algorithm and Extensions*, Wiley, New York (1997).

7. Wei C. G., Tanner M. A., A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms, *J. Am. Stat. Ass.* 85: 699, 1990.

Appendix 1

Bartoszyński, R., Edler, L., Hanin, L., Kopp-Schneider, A., Pavlova,L., Tsodikov, A., Zorin, A., and Yakovlev, A. Modeling cancer detection: Tumor size as a source of information on unobservable stages of carcinogenesis, *Mathematical Biosciences*, 2001, vol. 171, pp. 113-142.

The paper is attached.

Appendix 2

Boucher, K.M. and Kerber, R.A. The shape of the hazard function for cancer incidence, *Mathematical and Computer Modelling* 33: 1361-1376, 2001.

The paper is attached.

Appendix 3

Boucher, K.M. and Kerber, R.A. Measures of Familial Aggregation as Predictors of Breast Cancer Risk, *Journal of Epidemiology and Biostatistics* 6(5): 377-385, 2001.

The paper is attached.

Appendix 4

Hanin, L.G., Tsodikov, A.D., and Yakovlev, A.Y. Optimal schedules of cancer surveillance and tumor size at detection, *Mathematical and Computer Modelling* 33: 1419-1430, 2001.

The paper is attached.

PERGAMON Mathematical and Computer Modelling 33 (2001) 1419–1430

MATHEMATICAL AND COMPUTER MODELLING

www.elsevier.nl/locate/mcm

Optimal Schedules of Cancer Surveillance and Tumor Size at Detection

L. G. HANIN
Department of Mathematics, Idaho State University
Pocatello, ID 83209-8085, U.S.A.

A. D. TSODIKOV AND A. YU. YAKOVLEV
Huntsman Cancer Institute and Department of Oncological Sciences
University of Utah, 2000 East North Campus Drive
Salt Lake City, UT 84112-5550, U.S.A.

Dedicated to the memory of Robert Bartoszyński

Abstract—The paper explores methodological and mathematical aspects of a new approach to constructing optimal schedules of cancer screening. This approach consists of systematic use of tumor size at detection, combining stochastic models of tumor latency, tumor growth and tumor detection, and employing a new biologically natural screening efficiency criterion defined as the Kantorovich distance between the tumor size at spontaneous detection in the absence of screening and the tumor size at detection when both spontaneous and screening based mechanisms are in place. An explicit formula for the efficiency functional is obtained. Sample calculations suggest that in the case of exponential tumor growth, the optimal screening schedules with a fixed number of exams have a trend to uniformity.

Keywords—Screening, Optimal schedules, Tumor onset, Tumor size, Carcinogenesis models.

1. INTRODUCTION

Because of the significant cancer incidence and progress of tumor detection technology, cancer surveillance and screening are becoming increasingly important and costly public health problems. It is clear that appropriate mathematical methods are indispensable for a more effective management of the caseload through designing optimal surveillance strategies. Interest in exploring this avenue has quickened in the past few years [1–17].

The present work discusses methodological aspects of a new approach to optimization of cancer screening allowing for cancer detection at the earliest stages of tumor development. This makes the chances of tumor cure more favorable, reducing the probability of tumor recurrence. The

This work was supported by Grant DAMD17-98-1-8256 awarded by the U.S. Army Medical Research Acquisition Activity. The contents of this paper are solely the responsibility of the authors and do not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. The research of L. G. Hanin was also supported by Grant No. 807 from the Faculty Research Committee, Idaho State University, Pocatello, ID. We are very grateful to the reviewers whose comments have led to substantial improvements in the manuscript.

0895-7177/01/$ - see front matter Typeset by AMS-TEX
PII: S0895-7177(00)00325-3

problem of optimal cancer surveillance is set up as a search for optimal scheduling of screens subject to certain constraints on the number and timing of medical exams. Problems of a similar nature have already been addressed in the literature. Yakovlev and Tsodikov [1] have developed methods for constructing optimal surveillance strategies based on the minimum delay time criterion, given that the total number of examinations is fixed; see also [2]. They used dynamic programming methodology to solve the associated optimization problem. Their results show that this approach holds much promise for further practical use. As one example, the current practice for the breast cancer post-treatment surveillance at Curie Institute (Paris, France) is to examine the patients once per semester for the first four years, once per year for the next six years, and once every two years for the remaining period. For this strategy, the estimated false negative rate appears to be equal to 0.2 with the mean delay of the recurrence detection 4.1 months. Taking advantage of a previously proposed parametric model of tumor recurrence [3], the authors constructed the optimal strategy that provides a 33% reduction in the delay time, with the tests that comprise the optimal surveillance schedule tending to be more frequent when the hazard rate for the time to tumor is high. However, there are two weak points in this approach. First, the probability of tumor detection is assumed to be independent of the process of tumor regrowth. Second, estimation of the tumor onset time distribution is feasible only if a sample of diagnostic times produced by a discrete surveillance program with known false negative rate is available. The same applies equally to prediagnosis screening programs.

An alternative approach to the problem is to minimize the average cost of surveillance accounting for both examination costs and costs of late detection [1,4–15]. Since the two cost constituents are linked in the optimization procedure, the cost-utility approach makes it possible to search for both the optimal number of examinations and their sequence in time. However, the costs of late detection are usually very difficult to evaluate. For yet another optimization criterion based on the power of a statistical test for mortality rates, the reader is referred to [16].

Focusing our effort on possible medical rather than economic benefits, we propose to explore a new approach to the problem which is based on tumor size at detection. Tumor size is one of the most clinically significant characteristics of tumor maturity that determines largely the probability of both spontaneous and screening based tumor detection. This approach makes it possible to utilize data on tumor size at detection as an additional source of information on the natural history of the disease; some readily available epidemiologic data obtained from the control population in the absence of screening appear to be sufficient for estimation purposes. Another advantage of this approach is that it offers a natural way for incorporating the stage of tumor progression, where cancer detection normally occurs, into stochastic models of carcinogenesis. The proposed model of tumor progression accommodates a wide range of deterministic and stochastic laws of tumor growth.

As a measure of the effect of screening, we propose to use the difference between the expected tumor sizes at detection with and without screening, which coincides with the Kantorovich distance [18–21] between the distributions of the corresponding random variables. The structure of this distance allows for characterizing the net effect of screening, as compared to that of spontaneous detection.

Further advancements of the proposed approach to constructing optimal schedules of cancer screening will hopefully give answers to the following questions of major theoretical and practical importance.

1. Is the optimal efficiency of screening high enough to warrant its implementation?
2. What is the relation between the optimal screening schedules and their efficiencies for the criteria based on the tumor size and the expected time delay?
3. What are cancer specific patterns of optimal screening schedules?
4. What is the impact of hypothesized laws of tumor growth on the optimal screening efficiency and the pattern of the optimal examination schedules?

5. What are quantitative characteristics of the initiation, promotion, and progression stages for specific cancers?

The structure of the present paper is as follows. In Section 2, we describe some models of the natural history of cancer (including cancer latency and growth), screening schedules, and cancer detection. Here, we also formulate basic assumptions and introduce mathematical formalism. An explicit formula for the efficiency functional is derived in Section 3. Sample numerical calculations and analysis of their results are addressed in Section 4.

2. BASIC NOTIONS

2.1. Models of Carcinogenesis

In describing the natural history of cancer, the process of tumor development can be broken down into three stages. These stages are:

- formation of initiated cells,
- promotion of initiated cells resulting in appearance of the first malignant clonogenic cell, and
- subsequent growth and progression of malignant tumor.

The duration of each stage of carcinogenesis is thought of as a random variable (r.v.). In our sample calculations presented in Section 4, we use a two-parameter gamma family to specify the distribution of the length of the first two stages of carcinogenesis. However, more elaborate mechanistic models of carcinogenesis are available to describe the time to the event of malignant transformation. We provide two examples of such models.

The most widely accepted model of tumor latency is commonly referred to as the Moolgavkar-Venzon-Knudson (MVK) Model [22,23]. This Markovian two-stage model involves four parameters that refer to the rates of initiation of target stem cells (that is, formation of primary precancerous lesions), and rates of division, death or differentiation, and malignant transformation of initiated cells. It was first pointed out by Heidenreich [24] and subsequently by Hanin and Yakovlev [25] and Heidenreich *et al.* [26] that these four parameters are not jointly identifiable from time-to-tumor data. In the case of constant parameters, all triples of their identifiable combinations were described at length in [25]. In the latter case, the MVK model leads to the following explicit formula for the distribution of the total duration T of the first two stages, that is, of the time from the birth of an individual to the tumor onset [27,28],

$$\bar{F}_T(t) := \Pr(T > t) = \left[\frac{(a+b)e^{at}}{b + ae^{(a+b)t}}\right]^{\rho}, \qquad t \geq 0. \tag{1}$$

Here $a, b, \rho > 0$ are identifiable parameters of the model, $\bar{F}_T := 1 - F_T$ is the survivor function of the r.v. T, and F_T is the cumulative distribution function (c.d.f.) of the r.v. T.

Another model of carcinogenesis was proposed by Yakovlev and Polig in [29]. According to this model, the hazard function ϕ of the time T of tumor latency, which is related to the survivor function by

$$\bar{F}_T(t) = e^{-\int_0^t \phi(s)\,ds}, \qquad t \geq 0, \tag{2}$$

is of the form

$$\phi(s) = \theta_1 e^{-\theta_2 \int_0^s h(u)\,du} \int_0^s h(u) f(s-u)\,du, \qquad s \geq 0, \tag{3}$$

where h is a given time-dependent rate of external exposure, f is the probability density function (p.d.f.) of the tumor promotion time, and θ_1,θ_2 are positive constants. The key feature of the Yakovlev-Polig model is that it allows for the process of cell death to compete with the process of tumor promotion. Two particular cases of the model referring to spontaneous and induced carcinogenesis were employed in [30] and [31] to study the distribution of tumor size under a

threshold type mechanism of tumor detection. Recently, Hanin and Boucher [32] found conditions under which the parameters f, θ_1, θ_2 of the model given by (3) are identifiable from time-to-tumor observations. Specifically, a general necessary condition for identifiability of model (3) is given by the following theorem.

THEOREM 1. *Suppose that the function h satisfies $\int_0^\infty h(t)\,dt < \infty$ and that, for some $T > 0$, $h(t) = 0$ for $t > T$. If the model is identifiable in a family $\mathcal{F}$, then*

$$F(T) > 0, \qquad \text{for all } F \in \mathcal{F}.$$

DEFINITION. *A family $\mathcal{F}$ of absolutely continuous probability distributions on $\mathbf{R}_+$ is said to be graduated if for every two distinct p.d.f.s f, $\tilde{f} \in \mathcal{F}$ and for every constant $A > 0$, there is a number $\tau > 0$ (which may depend on f,$\tilde{f}$ and A) such that either $Af(t) \geq \tilde{f}(t)$ for all $t \geq \tau$, or $Af(t) \leq \tilde{f}(t)$ for all $t \geq \tau$.*

The following result generalizes Theorem 1 in the case of graduated families.

THEOREM 2. *Suppose that h is bounded, supported on $[0,T]$ for some $T > 0$, and positive almost everywhere on $[0,T]$. Then the model is identifiable in a graduated family $\mathcal{F}$ if and only if $F(T) > 0$ for all $F \in \mathcal{F}$.*

2.2. Tumor Growth

The following general functional form is assumed for the tumor size (the number of cells in a tumor) S:

$$S(w) = f_\theta(w), \tag{4}$$

where w is the time from the moment of the onset of cancer, and θ is a parameter which may be scalar or vector, deterministic or random. It is assumed that, for every θ, f_θ is a strictly monotonously increasing absolutely continuous function such that $f_\theta(0) = 1$. For a given θ, denote by g_θ the inverse function for f_θ, and set

$$\Phi_\theta(w) := \int_0^w f_\theta(u)\,du.$$

Specific laws of tumor growth of primary interest are listed below.

(1) Deterministic exponential growth; in this case, $S(w) = e^{\lambda w}$, where $\lambda > 0$ is a constant growth rate; see [33] for substantiation.
(2) Exponential growth with λ thought of as a gamma distributed r.v. [34].
(3) The Gompertz law

$$S(w) = e^{A\left(1-e^{-Bw}\right)},$$

with constant parameters $A, B > 0$.

2.3. Screening Schedules

The sequence of moments of time assigned for medical exams for a specific cancer and counted from the birth of a patient will be called a *screening schedule*. Let $\mathcal{T}$ be the set of all possible screening schedules $\tau = \{\tau_1 < \tau_2 < \cdots < \tau_n\}$. The set $\mathcal{T}$ may be subject to (some of) the following restrictions:

(a) $n \leq n_0$, where n_0 is an upper bound for the number of exams;
(b) $\tau_1 \geq m$ and $\tau_n \leq M$, where m and M are the earliest and the latest times for the first and the last exams, respectively;
(c) $\tau_{i+1} - \tau_i \geq h > 0$ for all $i = 1, 2, \ldots, n-1$ (this condition suggests a lower bound h for the minimal duration between any two successive exams).

Other restrictions on the moments of exams can also be accommodated. In the language of control theory, the set $\mathcal{T}$ is referred to as the set of admissible schedules.

2.4. Tumor Detection

We distinguish between *spontaneous* and *screening based* tumor detections. The first occurs in the absence of or concurrently with screening and is thought of as a continuous process. In contrast to this, screening based detection is an instantaneous event that may occur only at the moments of the prescribed medical exams and is therefore a discrete process. When both types of detection are present, they can be viewed as competing risks.

Numerous attempts have been made to relate the probability of detecting a tumor to its size [33–37]. Following [37], we assume that the rate r_0 of spontaneous tumor detection is proportional to the current tumor size

$$r_0 = \alpha_0 S, \tag{5}$$

where α_0 is a positive constant.

Let r.v.s W_0 and W_1 denote the times of spontaneous and screening based detections, counted from the moment of cancer onset, respectively. Then, for the moment W of combined detection, when both detection mechanisms are in place, we have $W = \min(W_0, W_1)$. Denote by

$$N_0 = f_\theta(W_0) \qquad \text{and} \qquad N = f_\theta(W) \tag{6}$$

the corresponding tumor sizes at spontaneous and combined detection.

Keeping in mind relation (2) between the survivor function of an absolutely continuous nonnegative r.v. and its hazard rate, we derive from (5) that, in the case of nonrandom parameter θ,

$$\bar{F}_{W_0}(w) = e^{-\int_0^w r_0(u)\,du} = e^{-\alpha_0 \int_0^w f_\theta(u)\,du} = e^{-\alpha_0 \Phi_\theta(w)}. \tag{7}$$

Therefore,

$$\bar{F}_{N_0}(n) = \bar{F}_{W_0}(g_\theta(n)) = e^{-\alpha_0 \Phi_\theta(g_\theta(n))},$$

and hence,

$$EN_0 = 1 + \int_1^\infty \bar{F}_{N_0}(n)\,dn = 1 + \int_1^\infty e^{-\alpha_0 \Phi_\theta(g_\theta(n))}\,dn = 1 + \int_0^\infty e^{-\alpha_0 \Phi_\theta(u)} f'_\theta(u)\,du. \tag{8}$$

If θ is a r.v., then an additional integration in (8) with respect to the distribution of θ is required.

In particular, for nonrandom exponential tumor growth with rate λ, we have

$$\bar{F}_{W_0}(w) = e^{-(\alpha_0/\lambda)\left(e^{\lambda w}-1\right)}, \qquad w \geq 0, \tag{9}$$

$$\bar{F}_{N_0}(n) = e^{-(\alpha_0/\lambda)(n-1)}, \qquad n \geq 1, \tag{10}$$

and

$$EN_0 = 1 + \frac{\lambda}{\alpha_0}. \tag{11}$$

Equation (10) suggests that in this case the r.v. N_0 has a translated exponential distribution with parameter α_0/λ. If λ is a r.v. which is gamma distributed with parameters μ, ν, then it follows from (11) that

$$EN_0 = 1 + \frac{\mu}{\alpha_0 \nu}.$$

We now specify the distribution of the r.v. W_1. Recall that W_1 is the time of screening based detection (in the absence of spontaneous detection) counted from the moment of appearance of the first malignant clonogenic cell. Indeed, the distribution of W_1 depends on the selected screening schedule $\tau = \{\tau_1 < \tau_2 < \cdots < \tau_n\}$. For the sake of convenience, set $\tau_0 := 0$ and $\tau_{n+1} := \infty$. It suffices to define, for every $t \geq 0$, the conditional distribution of W_1 given that $T = t$.

Let $\tau_i \le t < \tau_{i+1}$, $0 \le i \le n$. For $0 \le i \le n-1$ and $i+1 \le k \le n$, define the probability $p_t(k) := \Pr(W_1 = \tau_k - t \mid T = t)$ of tumor detection at the k^{th} screen given the cancer onset at moment t, and by $p_t(\infty) = 1 - \sum_{k=i+1}^{n} p_t(k)$ the corresponding conditional probability that the tumor is not detected by screening.

We introduce a discrete analogue of the hazard rate for the screening based detection by

$$\mu_t = \sum_{k=i+1}^{n} r_t(k)\delta_{\tau_k - t}, \tag{12}$$

where δ_x stands for the Dirac measure at x, and the sum over the empty set of indices is set, as usual, to be zero. By definition, the discrete measure μ_t is related to the conditional survivor function of W_1 given that $T = t$ through the equation

$$\bar{F}_{W_1|T=t}(w) = e^{-\int_0^w d\mu_t(u)}, \qquad w \ge 0, \tag{13}$$

compare with (2). It follows from (12) and (13) that

$$\sum_{j=k+1}^{n} p_t(j) + p_t(\infty) = \left[\sum_{j=k}^{n} p_t(j) + p_t(\infty)\right] e^{-r_t(k)}$$

or, equivalently, that

$$1 - \sum_{j=i+1}^{k} p_t(j) = \left[1 - \sum_{j=i+1}^{k-1} p_t(j)\right] e^{-r_t(k)}. \tag{14}$$

For $k = i+1$, we find from (14) that

$$1 - p_t(i+1) = e^{-r_t(i+1)}. \tag{15}$$

More generally, iterating this argument we obtain that

$$p_t(k) = e^{-\sum_{j=i+1}^{k-1} r_t(j)} \left[1 - e^{-r_t(k)}\right], \qquad i+1 \le k \le n.$$

Observe that this holds true for all $k = 1, \ldots, n$, if we set $p_t(k) = r_t(k) = 0$ for $1 \le k \le i$.

Similar to (5), we are assuming that the discrete rate of screening based detection is proportional to the current tumor size

$$r_t(k) = \alpha S(\tau_k - t), \qquad i+1 \le k \le n, \tag{16}$$

with some constant $\alpha > 0$. Combining (13), (12), and (16) with (4), we find that, given any t such that $\tau_i \le t < \tau_{i+1}$, $0 \le i \le n-1$,

$$\bar{F}_{W_1|T=t}(w) = e^{-\alpha \sum_{k=i+1}^{j} f_0(\tau_k - t)}, \qquad \text{where } \tau_j - t \le w < \tau_{j+1} - t, \quad i+1 \le j \le n. \tag{17}$$

Consider the case of one exam occurring at a moment τ with the detection probability $p = p(t,\tau)$ and the discrete detection rate $r = r(t,\tau)$. Then by (15), $1 - p = e^{-r}$. If the probability p is small, then the rate r is approximately equal to p. In particular, under assumption (16), the probability of tumor detection is approximately proportional to the current tumor size $p \simeq \alpha S(\tau - t)$. Klein and Bartoszyński [34] proceeded in their study of breast cancer from a more general assumption that the probability of tumor detection is proportional to some power of the tumor size. Their estimate of this power leads, however, to a value which is very close to 1.

3. FORMULA FOR THE SCREENING EFFICIENCY FUNCTIONAL

We proceed from the following two biologically natural assumptions.

1. The r.v.s W_0 and T are independent.
2. For every $t \geq 0$, the r.v.s W_1 and W_0 are conditionally independent given that $T = t$.

The first assumption claims that the moment of spontaneous tumor detection measured from the appearance of the first malignant clonogenic cell is independent of the prior duration of tumor latency. The second assumption reflects a technological (or instrumental) nature of both detection processes. It states that, given the moment of cancer onset, the two times W_0 and W_1, at which competing events of the spontaneous and screening based tumor detection may occur, are independent. This statement immediately follows from the assumption that both detection processes are completely determined by the current tumor size as a deterministic function of time.

For an admissible screening schedule $\tau \in \mathcal{T}$, we define the efficiency functional as the Kantorovich distance $d_K(N_0, N; \tau)$ (see [18,20,21]) between the tumor sizes N_0 and N at spontaneous and combined detection. This quantity serves as a clinically natural measure of the gain resulting from screening. It is well known [19,20] that

$$d(N, N_0; \tau) = \int_1^\infty \left| \bar{F}_{N_0}(n) - \bar{F}_N(n) \right| dn. \tag{18}$$

It follows from (7), inequality $W_0 \geq W$, and monotonicity of the function f_θ that the r.v. N_0 stochastically dominates the r.v. N : $\bar{F}_{N_0} \geq \bar{F}_N$. This leads to the following alternative expression for the efficiency functional:

$$d(N, N_0; \tau) = \int_1^\infty \bar{F}_{N_0}(n)\, dn - \int_1^\infty \bar{F}_N(n)\, dn = EN_0 - EN, \tag{19}$$

where E stands for the expectation.

Suppose that parameter θ is nonrandom. We set $n = f_\theta(w)$ and condition upon the r.v. T in (18) to obtain

$$\begin{aligned} d(N, N_0; \tau) &= \int_0^\infty \left| \bar{F}_{W_0}(w) - \bar{F}_W(w) \right| f'_\theta(w)\, dw \\ &= \int_0^\infty \int_0^\infty \left| \bar{F}_{W_0}(w) - \bar{F}_{W|T=t}(w) \right| f'_\theta(w)\, dw\, dF_T(t), \end{aligned}$$

where $\bar{F}_{W|T=t}$ is the conditional survivor function of the r.v. W given that $T = t$. Since $W = \min(W_0, W_1)$, it follows from our Assumptions 1 and 2 that

$$\bar{F}_{W_0} - \bar{F}_{W|T=t} = \bar{F}_{W_0} - \bar{F}_{W_0}\bar{F}_{W_1|T=t} = \bar{F}_{W_0} F_{W_1|T=t}.$$

Therefore,

$$d(N, N_0; \tau) = \int_0^\infty \int_0^\infty F_{W_1|T=t}(w) \bar{F}_{W_0}(w) f'_\theta(w)\, dw\, dF_T(t). \tag{20}$$

Observe that if $T = t$, where $\tau_i \leq t < \tau_{i+1}$, $0 \leq i \leq n$, then the only possible values of the r.v. W_1 are $\tau_{i+1} - t, \ldots, \tau_{n+1} - t$. More specifically, $W_1 = \tau_j - t$, $i+1 \leq j \leq n$, if the j^{th} exam detected a tumor, and $W_1 = \tau_{n+1} - t = \infty$ if the tumor was not detected in the course of screening. Therefore, if $t \geq \tau_n$ or $\tau_i \leq t < \tau_{i+1}$, $0 \leq i \leq n-1$, and $0 \leq w < \tau_{i+1} - t$, then $F_{W_1|T=t}(w) = 0$. This allows us to rewrite (20) in the form

$$d(N, N_0; \tau) = \sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} \sum_{j=i+1}^{n} \int_{\tau_j - t}^{\tau_{j+1} - t} F_{W_1|T=t}(w) \bar{F}_{W_0}(w) f'_\theta(w)\, dw\, dF_T(t).$$

We now recall the explicit expression (17) derived above for the function $\bar{F}_{W_1|T=t}$, and denote

$$G_\theta(x) := \int_x^\infty \bar{F}_{W_0}(w) f'_\theta(w)\, dw, \qquad x \geq 0,$$

to obtain finally

$$\begin{aligned} d(N, N_0; \tau) &= \sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} \sum_{j=i+1}^{n} \left[1 - e^{-\alpha \sum_{k=i+1}^{j} f_\theta(\tau_k - t)}\right] [G_\theta(\tau_j - t) - G_\theta(\tau_{j+1} - t)]\, dF_T(t) \\ &= \sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} \sum_{j=i+1}^{n} e^{-\alpha \sum_{k=i+1}^{j-1} f_\theta(\tau_k - t)} \left[1 - e^{-\alpha f_\theta(\tau_j - t)}\right] G(\tau_j - t)\, dF_T(t). \end{aligned} \tag{21}$$

In the case when parameter θ is random, the right-hand side of (21) should be integrated additionally with respect to the distribution of θ.

If, in particular, $f_\theta(w) = e^{\lambda w}$ with a constant rate λ, then invoking (9) we find easily that

$$G_\theta(x) = \frac{\lambda}{\alpha_0} e^{-(\alpha_0/\lambda)\left(e^{\lambda x} - 1\right)}, \qquad x \geq 0.$$

In this case, the efficiency functional (21) takes on the form

$$\begin{aligned} &d(N, N_0; \tau) \\ &= \frac{\lambda}{\alpha_0} \sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} \sum_{j=i+1}^{n} e^{-\alpha \sum_{k=i+1}^{j-1} e^{\lambda(\tau_k - t)}} \left[1 - e^{-\alpha e^{\lambda(\tau_j - t)}}\right] e^{-(\alpha_0/\lambda)\left(e^{\lambda(\tau_j - t)} - 1\right)} dF_T(t). \end{aligned} \tag{22}$$

Observe also that (19) implies

$$EN = EN_0 - d(N, N_0; \tau).$$

This allows for an explicit calculation of the expected tumor size at combined detection on the basis of formulas (8) and (21).

The problem

$$d(N, N_0; \tau) \to \max, \qquad \tau \in \mathcal{T}, \tag{23}$$

can be solved by exhaustive search with some simplification arising from the special form of the dependence of the functional (21) on τ. A question of practical importance is what are the values of the number n of exams for which the problem (23) is computationally feasible. We will conclude this paper, which deals primarily with methodological and mathematical aspects of the problem of optimization of cancer surveillance, with some sample calculations with prescribed values of model parameters.

4. NUMERICAL EXPERIMENTS

It was assumed that the time T to tumor onset is gamma distributed with the mean $\mu = 50$ years and the standard deviation $\sigma = 20$ years. The graph of the c.d.f. F_T is shown in Figure 1. The law of tumor growth was taken to be deterministic exponential with the rate $\lambda = 1.6$ years^{-1}, which corresponds to the tumor size doubling time of approximately 5.2 months. The rate of spontaneous tumor detection was assumed to be $\alpha_0 = 0.03$. This value has no relevance to any actual data and serves only a purpose of illustration. The graph of the survivor function $\bar{F}_{W_0}$ given by equation (9) is presented in Figure 2. The effect of one exam occurring after tumor onset with the screening based tumor detection rate $\alpha = 0.1$ is shown in Figure 3 featuring the survivor function of the time W to combined tumor detection.

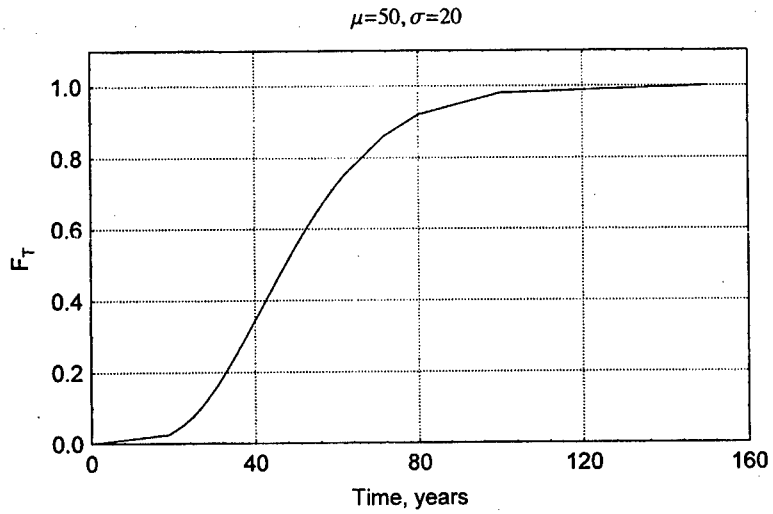


Figure 1. The cumulative distribution function for the time to tumor onset.

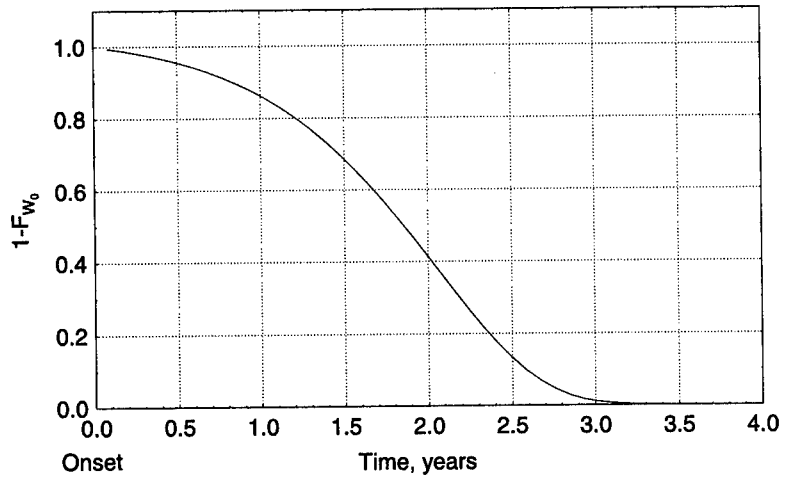


Figure 2. The survivor function for the time to spontaneous detection.

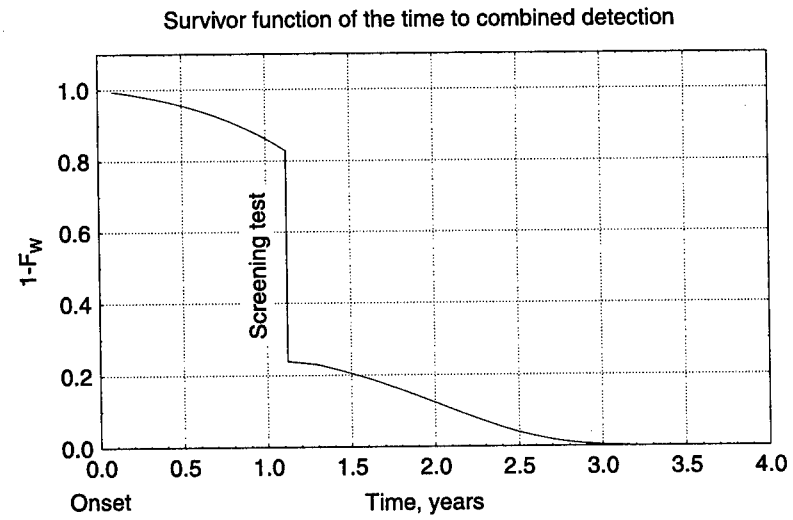


Figure 3. The survivor function for the time to combined detection ($\alpha = 0.1$).

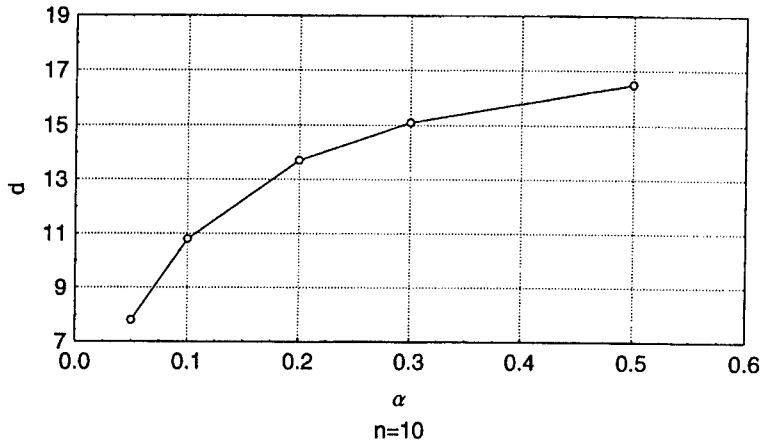


Figure 4. Optimal screening efficiency as a function of the parameter α given a fixed number ($n = 10$) of examinations and $\sigma = 20$ years.

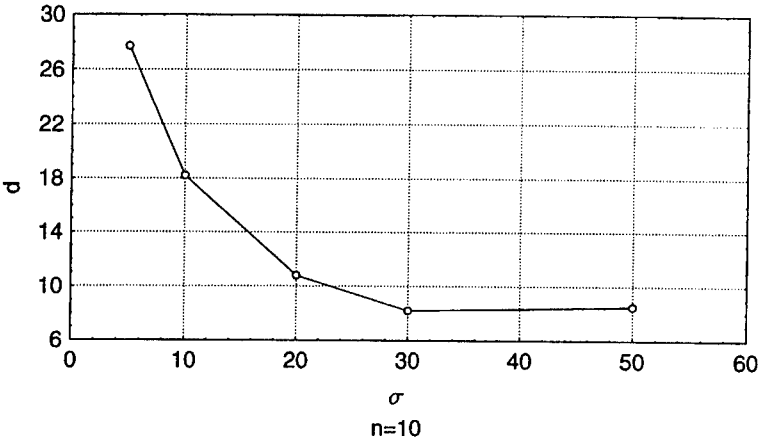


Figure 5. Optimal screening efficiency as a function of the standard deviation, σ, of the onset time ($n = 10$, $\alpha = 0.1$).

Table 1. Optimal screening schedules.

$n = 20$											
σ	α	$\Delta_l = \tau_1$	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6	Δ_7	Δ_8	Δ_9	Δ_{10}
20	0.1	27.75	2.25	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.25
		Δ_{11}	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}	Δ_{18}	Δ_{19}	Δ_{20}
		2.25	2.25	2.25	2.25	2.25	2.25	2.25	2.25	2.25	2.25
$n = 10$											
σ	α	$\Delta_l = \tau_1$	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6	Δ_7	Δ_8	Δ_9	Δ_{10}
20	0.1	35.00	2.50	2.50	2.50	2.25	2.50	2.50	2.50	2.50	2.50
20	0.3	35.00	2.50	2.50	2.50	2.50	2.50	2.50	2.50	2.50	2.50
20	0.5	34.50	2.75	2.75	2.50	2.25	2.50	2.50	2.50	2.75	2.75
10	0.1	40.75	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.25
30	0.1	27.25	2.50	2.75	2.50	2.50	2.50	2.50	2.50	2.50	2.50

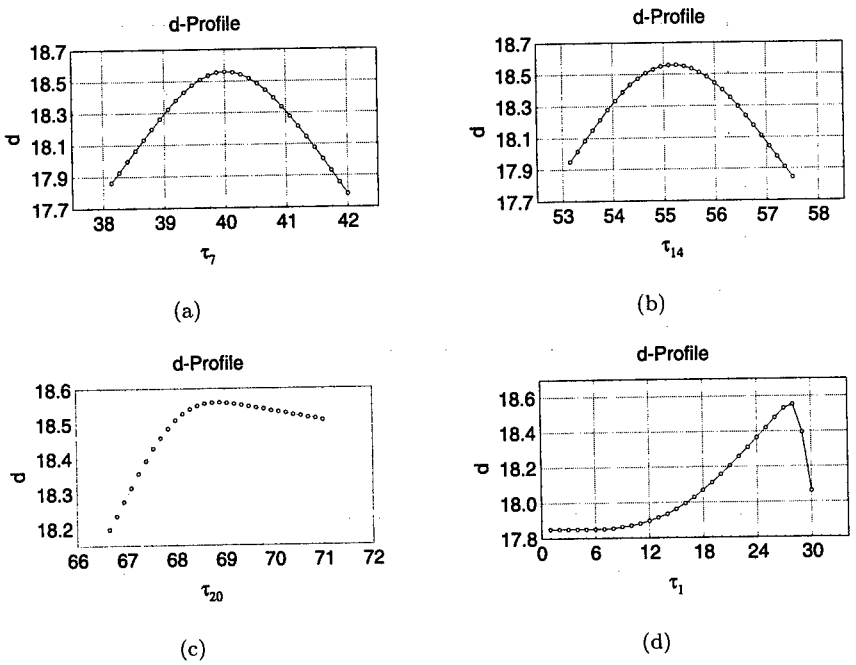


Figure 6. Profiles of the efficiency functional ($n = 20$, $\sigma = 20$ years, $\alpha = 0.1$).

The search for optimal screening schedules and optimal screening efficiencies was conducted for a fixed number n of screens with no restriction on the moments of exams and for various values of α. The method of optimization was the exhaustive search with the step 0.25 years. Parameter values $\mu = 50$ years, $\lambda = 1.6$ years^{-1}, and $\alpha_0 = 0.03$ were fixed throughout the calculations. For $n = 10$, plots of the rescaled optimal screening efficiency d with $\sigma = 20$ years versus α and, for $\alpha = 0.1$, versus σ are shown in Figures 4 and 5, respectively. As it could be expected, d increases with increasing α and decreases with increasing σ.

The results of our search for optimal screening schedules with $n = 10, 20$ and with several values of σ and α are given in Table 1. For the reader's convenience, screening schedules are represented by the intervals $\Delta_i := \tau_i - \tau_{i-1}$, $i = 1, \ldots, n$, between two successive exams. For all cases explored, optimal screening schedules are uniform or very close to such.

As a test for optimality of a screening schedule, profiles of the efficiency functional (22), with $n - 1$ moments of exams fixed at the optimal values and the remaining one varying between the two fixed neighboring moments of exams, were computed. For $n = 20$ and $\alpha = 0.1$, these profiles are given in Figure 6. For the moment τ_1, a clear cut maximum was observed (see Figure 6d), while for τ_{20} the maximum is more flat (see Figure 6c). All intermediate moments of exams $\tau_2, \ldots, \tau_{19}$ demonstrated a well-pronounced parabolic maximum (see Figures 6a and 6b).

This study shows that mathematical and computational problems of optimal cancer surveillance are tractable within the framework of the proposed model.

REFERENCES

1. A.Yu. Yakovlev and A.D. Tsodikov, *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific, Singapore, (1996).
2. A.D. Tsodikov, B. Asselain, A. Fourquet, T. Hoang and A.Yu. Yakovlev, Discrete strategies of cancer post-treatment surveillance. Estimation and optimization problems, *Biometrics* **51**, 437–447, (1995).
3. A.Yu. Yakovlev, B. Asselain, V.-J. Bardou, A. Fourquet, T. Hoang, A. Rochefodiere and A.D. Tsodikov, A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer, In *Biométrie et Analyse de Données Spatio-Temporelles, Vol. 12*, (Edited by B. Asselain *et al.*), pp. 66–82, Société Française de Biométrie, ENSA, Rennes, (1993).
4. R. Kirch and M. Klein, Surveillance schedules for medical examinations, *Management Science* **20**, 1403–1409, (1974).

5. A.K. Shahani and D.M. Crease, Towards models of screening for early detection of disease, *Adv. Appl. Prob.* **9**, 665–680, (1977).
6. D.M. Eddy, A mathematical model for timing repeated medical tests, *Medical Decision Making* **3**, 34–62, (1983).
7. A.D. Tsodikov and A.Yu. Yakovlev, On the optimal policies of cancer screening, *Math. Biosci.* **107**, 21–45, (1991).
8. A.D. Tsodikov, A.Yu. Yakovlev and L. Petukhov, Some approaches to screening optimization, In *Statistique des Processus en Milieu Médical*, (Edited by B. Bru *et al.*), pp. 1–48, Université Paris V, Paris, France, (1991).
9. A.D. Tsodikov, Screening under uncertainty. Games approach, *Syst. Anal. Model. Simul.* **9**, 259–262, (1992).
10. G. Parmigiani, On optimal screening schedules, *Biometric Bulletin* **8** (3), 21, (1991).
11. G. Parmigiani, Optimal scheduling of fallible inspections, DP 92-A38, ISDS, Duke University, (1992).
12. G. Parmigiani, On optimal screening ages, *J. Amer. Statist. Assoc.* **88**, 622–628, (1993).
13. G. Parmigiani, Timing medical examinations via intensity functions, *Biometrika* **84** (4), 803–816, (1997).
14. G. Parmigiani and M.S. Kamlet, A cost-utility analysis of alternative strategies in screening for breast cancer, In *Bayesian Statistics in Science and Technology: Case Studies*, (Edited by C. Gatsonis *et al.*), pp. 390–402, Springer, New York, (1993).
15. M. Zelen, Optimal scheduling of examinations for the early detection of disease, *Biometrika* **80**, 279–293, (1993).
16. P. Hu and M. Zelen, Planning clinical trials to evaluate early detection programmes, *Biometrika* **84** (4), 817–829, (1997).
17. S.J. Lee and M. Zelen, Scheduling periodic examinations for the early detection of disease: Applications to breast cancer, *J. Amer. Statist. Assoc.* **93**, 1271–1281, (1998).
18. L.V. Kantorovich and G.P. Akilov, *Functional Analysis*, 2nd Edition, Pergamon Press, New York, (1982).
19. S.S. Vallander, Calculation of the Wasserstein distance between probability distributions on the line, *Theory Prob. Appl.* **18**, 784–786, (1973).
20. S.T. Rachev and R.M. Short, Duality theorems for Kantorovich-Rubinstein and Wasserstein functionals, *Dissertationes Math.* **299**, (1990).
21. L.G. Hanin, Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces, *Proc. Amer. Math. Soc.* **115** (2), 345–352, (1992).
22. S.H. Moolgavkar and D.J. Venzon, Two event model for carcinogenesis: Incidence curves for childhood and adult tumors, *Math. Biosci.* **47**, 55–77, (1979).
23. S.H. Moolgavkar and A.G. Knudson, Mutation and cancer: A model for human carcinogenesis, *J. Natl. Cancer Inst.* **66**, 1037–1052, (1981).
24. W.F. Heidenreich, On the parameters of the clonal expansion model, *Radiat. Envir. Biophys.* **35**, 127–129, (1996).
25. L.G. Hanin and A.Yu. Yakovlev, A nonidentifiability aspect of the two-stage model of carcinogenesis, *Risk Analysis* **16** (5), 711–715, (1996).
26. W.F. Heidenreich, E.G. Luebeck and S.H. Moolgavkar, Some properties of the hazard function of the two-mutation clonal expansion model, *Risk Analysis* **17**, 391–399, (1997).
27. A. Kopp-Schneider, C.J. Portier and C.D. Sherman, The exact formula for tumor incidence in the two-stage model, *Risk Analysis* **14**, 1079–1080, (1994).
28. Q. Zheng, On the exact hazard and survival functions of the MVK stochastic carcinogenesis model, *Risk Analysis* **14**, 1081–1084, (1994).
29. A.Yu. Yakovlev and E. Polig, A diversity of responses displayed by a stochastic model of carcinogenesis allowing for cell death, *Math. Biosci.* **132**, 1–33, (1996).
30. A.Yu. Yakovlev, L.G. Hanin, S.T. Rachev and A.D. Tsodikov, Distribution of tumor size at detection and its limiting form, *Proc. Natl. Acad. Sci. USA* **93**, 6671–6675, (1996).
31. L.G. Hanin, S.T. Rachev, A.D. Tsodikov and A.Yu. Yakovlev, A stochastic model of carcinogenesis and tumor size at detection, *Adv. Appl. Prob.* **29**, 607–628, (1997).
32. L.G. Hanin and K.M. Boucher, Identifiability of parameters in the Yakovlev-Polig model of carcinogenesis, *Math. Biosci.* **160**, 1–24, (1999).
33. R. Bartoszyński, A modeling approach to metastatic progression of cancer, In *Cancer Modeling*, (Edited by J.R. Thompson and B.W. Brown), pp. 237–267, Marcel Dekker, New York, (1987).
34. M. Klein and R. Bartoszyński, Estimation of growth and metastatic rates of primary breast cancer, In *Mathematical Population Dynamics*, (Edited by O. Arino *et al.*), pp. 397–412, Marcel Dekker, New York, (1991).
35. N.E. Atkinson, R. Bartoszyński, B.W. Brown and J.R. Thompson, On estimating the growth function of tumors, *Math. Biosci.* **67**, 145–166, (1983).
36. N.E. Atkinson, B.W. Brown and J.R. Thompson, On the lack of concordance between primary and secondary tumor growth rates, *J. Natl. Cancer Inst.* **78**, 425–435, (1987).
37. B.W. Brown, N.E. Atkinson, R. Bartoszyński and E.D. Montague, Estimation of human tumor growth rate from distribution of tumor size at detection, *J. Natl. Cancer Inst.* **72**, 31–38, (1984).

ELSEVIER

Mathematical Biosciences 171 (2001) 113–142

Mathematical Biosciences

www.elsevier.com/locate/mbs

Modeling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis

Robert Bartoszyński [a], Lutz Edler [b], Leonid Hanin [c], Annette Kopp-Schneider [b], Lyudmila Pavlova [d], Alexander Tsodikov [e], Alexander Zorin [e,f], Andrej Yu. Yakovlev [e,*]

[a] *Department of Statistics, 141 Cockins Hall, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA*
[b] *Biostatistics Unit, The German Cancer Research Center, D-69009 Heidelberg, Germany*
[c] *Department of Mathematics, Idaho State University, Pocatello, ID 83209-8085, USA*
[d] *Department of Applied Mathematics, St. Petersburg State Technical University, 29 Polytechnicheskaya Street, St. Petersburg 195251, Russia*
[e] *Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112-5550, USA*
[f] *The Central Research Institute of Radiology, 70/4, Leningradskaya Street, Pesochny-2, St. Petersburg 189646, Russia*

Received 31 August 2000; received in revised form 5 March 2001; accepted 21 March 2001

Abstract

This paper is concerned with modern approaches to mechanistic modeling of the process of cancer detection. Measurements of tumor size at diagnosis represent a valuable source of information to enrich statistical inference on the processes underlying tumor latency. One possible way of utilizing this information is to model cancer detection as a quantal response variable. In doing so, one relates the chance of detecting a tumor to its current size. We present various theoretical results emerging from this approach and illustrate their usefulness with numerical examples and analyses of epidemiological data. An alternative approach based on a threshold type mechanism of tumor detection is briefly described. © 2001 Elsevier Science Inc. All rights reserved.

Keywords: Cancer detection; Metastatic process; Quantal response; Stochastic models; Identifiability; Parametric estimation

* Corresponding author. Tel.: + 1-801 585 9544; fax: +1-801 585 5357.
E-mail address: yak@hci.utah.edu (A.Yu. Yakovlev).

PII: S0025-5564(01)00058-X

1. Introduction

Dr Robert Bartoszyński passed away on 17 January 1998. The biography of this exceptional individual will be published in a special issue of *Mathematical and Computer Modelling* [1] dedicated to his memory. In last months of his life he was developing new approaches to stochastic modeling of quantal response variables in carcinogenesis and metastatic progression of tumors. Unfortunately, this latest work remained unfinished. The originality and depth of his thought lead us to expect that new notable results would have emerged from his research endeavor. In this paper, Robert's friends and colleagues make an attempt to develop some of his most basic ideas. In doing so, we discuss a variety of issues and problems that arise in the quantitative description of the process of cancer detection, not only those of special interest to Robert in his last work. We are convinced that Robert would have done this differently and most probably in a much more elegant way. However, this is the best these authors could do to pay a tribute to one of the brightest scientists in the field of biomathematics and biostatistics.

Bartoszyński and co-workers [2–6] have developed a new avenue in stochastic modeling of cancer detection. The basic idea behind their approach is to relate the chance of detecting a tumor to its current size. This idea was also explored by Kimmel and Flehinger [7] in the context of the primary tumor size – metastasis relationship in solid cancers and by Hanin et al. [8] in an attempt to develop new approaches to optimal scheduling of cancer surveillance.

In the present communication, we address a wide spectrum of problems associated with stochastic modeling of cancer detection. In Sections 2 and 3, we give an introduction to the modeling techniques based on quantal response models. Marginal distributions of tumor size and age at detection as well as associated estimation problems are discussed in Sections 4 and 5; the joint distribution of the two random variables and their randomized counterpart is given in Section 8. Generally speaking, explicit formulas for the marginal distributions of tumor size and age of an individual at detection are not sufficient to utilize completely the information contained in the corresponding sample observations for estimation of parameters describing the natural history of the disease; one needs to know their joint distribution in order to develop pertinent methods for the maximum likelihood statistical inference.

In Sections 6 and 7, we explore some identifiability and stability properties of the proposed model of tumor detection. For the sake of completeness, an alternative model of a threshold type process of tumor detection is considered in Section 9. Section 10 explores similar ideas in relevance to metastatic processes which were also one of the major subjects for Dr. Bartoszyński's research in the latest period of his life.

We believe that the basic idea behind the modeling techniques presented in this paper deserves further exploration.

2. Quantal response variables

Let $Y(t)$, $t \geqslant 0$, be a stochastic process, and T be an absolutely continuous non-negative random variable (r.v.) defined on the same probability space and interpreted as time of occurrence of a certain event.

Definition 1. A random variable T is said to be quantal with respect to the stochastic process $Y(t)$ if there exists a non-negative risk function $r(y)$, $y \geqslant 0$, such that for all $t \geqslant 0$, $\Delta t > 0$, and all admissible y,

$$Pr\{T \in [t, t+\Delta t) | T > t \text{ and } Y(t) = y\} = r(y)\Delta t + o(\Delta t),$$

where $o(x)$ is a function such that $o(x)/x \to 0$ as $x \to 0+$.

The concept of quantal response variable was introduced and studied by Puri and Centuria [9]; its further analysis is due to Puri [10,11]. In fact, the definition in [9] postulated existence of a more general risk function $r(y, t)$; however, the present work is concerned with age-independent case where the conditional hazard function of the r.v. T depends on time only through the current value of the stochastic process $Y(t)$. Specifically, it is the important particular case $r(y) = \alpha y$ with constant $\alpha > 0$ that is the main subject matter of this paper.

The above concept appeared in [9] as an alternative to threshold models of biological effects. The latter modeling techniques are based on the assumption that a response of the organism to a given (pathological) process occurs as soon as the process exceeds some threshold level, which may be random or deterministic (see Section 9 for further discussion). Puri [10] (see also [9]) referred to a personal communication with LeCam, who apparently was the first to suggest a quantal response model in the context of host's response to microbial infection.

Whether a quantal mechanism, or a threshold mechanism is a more adequate description of the process of tumor detection is a question that may be impossible to answer in the present state of biological knowledge. The diversity and complexity of detection mechanisms remain to be clearly understood. In such situations, it is perhaps best to take a pragmatic approach, and choose that option which leads to more mathematically tractable formulations of biologically meaningful problems. The main difference between threshold and quantal mechanisms is that in the first case the event of interest always occurs at the so-called ladder point of the process $Y(t)$, that is, at a value higher than any previously attained. In contrast, a quantal event may occur at any value of the process $Y(t)$; it is only assumed that the likelihood of the event of interest depends on the process $Y(t)$. This means that the quantal response model is indeterministic in nature; its stochastic character remains even if the process $Y(t)$ is deterministic.

A more general quantal response model arises when the hazard function of the time T to the event of interest depends on the past *sample path* of the process $Y(t)$ or on functionals of the sample path. Handling quantal responses in this case involves conditioning on a sample path of the process $Y(t)$, deriving the desired formulas, and then 'unconditioning' the resultant expressions. The main difficulty lies, naturally, in the last step. In the opposite extreme case where $Y(t)$ is a deterministic strictly increasing function of time, various characteristics of r.v. T can be obtained through its hazard function $h_T(t) = r(Y(t))$.

3. Tumor detection as a quantal response event

When modeling time of tumor detection as a quantal response variable in accordance with Definition 1, the first problem is to determine the form of the risk function $r(y)$. The most natural approach is to relate the chance of detecting a tumor to its current size [2–8]. Although the rule 'the larger the tumor, the more likely it will become detected' appears unquestionably valid, the

real question is to what characteristic of tumor size is the 'risk' of detection proportional. For tumors easily accessible to palpation (e.g. skin cancer), the answer is obvious: it is the volume of a tumor that plays the role of $Y(t)$. The same seems to be true for breast cancer, although the reasons for this are much less clear. Indeed, the major technique of detecting breast tumors is mammography, where the tumor is observed as a projection on the plane. A priori, therefore, one could postulate that the process $Y(t)$ in breast cancer detection should be the surface of the projection, that is a quantity of the order of $[V(t)]^{2/3}$, where $V(t)$ is the volume of the tumor at time t. Klein and Bartoszyński [6] tested the relationship $Y(t) = [V(t)]^{\epsilon}$, and estimated ϵ, with $\epsilon = 1$ providing the best fit to breast cancer data. For other tumor sites, detection may be related to appearance and intensity of some symptoms, which may depend on cumulative effects of the presence of a tumor, so that $Y(t) = \int_0^t V(t-\tau)k(\tau)\,d\tau$ for some suitable functions or measures k. Another example of this kind is provided by molecular markers of tumor growth that have proven to be quite useful in clinical practice.

Suppose the growth of a tumor begins at time $t = 0$. Let $X(t)$ be the number of tumor cells at time $t \geqslant 0$ and $X(0) = n_0$. In the simplest case where $Y(t) = \alpha X(t)$, $\alpha > 0$, the following line of reasoning illustrates distinct advantages provided by the assumption that tumor detection is a quantal response event. Suppose that the initial number of tumor cells n_0 is non-random and the growth of a tumor obeys the postulates of the homogeneous pure birth process, also known as the Yule process [12], with a constant birth rate λ. Then for every fixed t the random variable $X(t)$ follows a negative binomial distribution, that is

$$Pr\{X(t) = n\} = \binom{n-1}{n-n_0} e^{-\lambda t n_0}(1 - e^{-\lambda t})^{n-n_0}, \quad n \geqslant n_0,$$

$$Pr\{X(t) = n\} = 0, \quad n < n_0.$$

It follows that the expectation and variance of tumor size at time t are equal to

$$E\{X(t)\} = X(0)e^{\lambda t} \quad \text{and} \quad \mathrm{Var}\{X(t)\} = X(0)e^{\lambda t}(e^{\lambda t} - 1), \tag{1}$$

respectively.

Let

$$Z(t) := X(t)e^{-\lambda t}.$$

Then, for $t, \tau \geqslant 0$ we have

$$\begin{aligned} E\{Z(t+\tau)|Z(t)\} &= E\{X(t+\tau)e^{-\lambda(t+\tau)}|Z(t)\} \\ &= e^{-\lambda(t+\tau)}E\{X(t+\tau)|X(t)\} \\ &= e^{-\lambda(t+\tau)}X(t)e^{\lambda\tau} \\ &= Z(t). \end{aligned}$$

This argument shows that $Z(t)$ is a martingale. Consequently, $X(t)e^{-\lambda t}$ converges almost surely to a random variable, say ξ, as $t \to \infty$ [13]. Under mild conditions, a similar asymptotic result holds for a more general model of the Bellman–Harris branching stochastic process [14]. It follows from formulas (1) that $E\{\xi\} = \mathrm{Var}\{\xi\} = n_0$. According to the widely accepted clonogenic concept of tumor development, the process of tumor growth begins with a single malignant cell, i.e. $n_0 = 1$, in

which case $E\{\xi\} = \mathrm{Var}\{\xi\} = 1$, and the random variable ξ has the unit exponential distribution [15]. Therefore, the contribution of initial stochastic fluctuations to the process $X(t)$ is expected to be relatively small if the time elapsed from the tumor onset is sufficiently large.

From the practical point of view, it would make sense to disregard the fluctuations, and treat $X(t)$ as a continuous increasing deterministic function of time. Consequently, the volume of a tumor at time t after the event that initiates the tumor growth (generates the first malignant cell) will be described as $ce^{t/\gamma}$, where $\gamma = 1/\lambda$ and $c = x(0)$ is the volume of a single tumor cell ($c \simeq 10^{-9}$ cm^3, see [6]).

There are at least two ways of generalizing this simple model. First, one may treat the parameter γ (or λ) as a random variable, thereby yielding various randomized counterparts of the basic model of exponential growth. The second possibility is to go beyond the pure birth process and proceed from more complex laws of tumor growth, like the Gompertz or logistic functions, to generate a richer family of models. In doing so, it is a good idea to construct an hierarchical (nested) family of models so that a particular (minimal) sub-family could be selected to provide a sufficiently accurate description of real data. Whether or not the above possibilities are feasible depends heavily on the type of data to be analyzed.

4. Tumor size at detection and its distribution

In this paper we confine ourselves to the process of spontaneous (without screening) tumor detection. This term refers to self-detection and incidental diagnoses resulting from individual medical exams that do not follow any fixed surveillance schedule. However, it is possible to extend the model of cancer detection to include discrete surveillance (screening) strategies or a combination of spontaneous and screening-based detection mechanisms [8]. It should be kept in mind that in most cases the event of spontaneous tumor detection is a process that takes a certain amount of time rather than an instantaneous event. This process normally consists of various medical exams that may or may not be triggered symptomatically. In what follows, spontaneous detection is thought of as occurring in the course of the exam that confirms inequivocally the tentative diagnosis, and the time of this exam is termed the time of spontaneous tumor detection.

Suppose that the number of tumor cells, $X(t)$, present in the tumor at time t (measured from the time of tumor onset, i.e. the event of malignant transformation of a premalignant initiated cell) is described by a deterministic growth function of time, that is,

$$X(t) = f_\theta(t), \tag{2}$$

where θ is a parameter which may be scalar or vector, deterministic or random. For simplicity, we assume that the parameter θ is non-random. It is also assumed that, for every θ, $f_\theta(t)$ is an absolutely continuous strictly monotonically increasing function such that $f_\theta(0) = 1$. For a given θ, denote by $\psi_\theta(t)$ the inverse function for $f_\theta(t)$, and set

$$\Phi_\theta(t) := \int_0^t f_\theta(u)\,\mathrm{d}u.$$

The next step is to assume that the rate r of spontaneous tumor detection is proportional to the current tumor size

$$r(t) = \alpha X(t) = \alpha f_0(t), \tag{3}$$

where α is a positive constant, compared with Definition 1. Let W be the time of spontaneous tumor detection measured from the time of tumor onset, and let S be the tumor size at detection. Then r is the hazard function of r.v. W. Therefore, the survival function, $\bar{G}_W(w) = Pr\{W > w\}$, for the random variable W is given by

$$\bar{G}_W(w) = \exp\left\{ -\int_0^w r(u)\,du \right\} = \exp\left\{ -\alpha \int_0^w f_0(u)\,du \right\} = e^{-\alpha\Phi_0(w)}. \tag{4}$$

In like manner, we derive the tumor size tail function, $\bar{F}_S(s) = Pr\{S > s\}$, of the random variable $S = f_0(W)$

$$\bar{F}_S(s) = \bar{G}_W[\psi_0(s)] = e^{-\alpha\Phi_0(\psi_0(s))}. \tag{5}$$

In the derivation of formula (5), it is assumed that $f(t) \to \infty$ as $t \to \infty$. If $f(t)$ has a finite limit, say d, then formula (5) is true for $0 \leqslant s < d$ and $\bar{F}_S(s) = 0$ for $s \geqslant d$.

The expected tumor size at detection is equal to

$$\mathbf{E}\{S\} = 1 + \int_1^\infty \bar{F}_S(s)\,ds = 1 + \int_0^\infty e^{-\alpha\Phi_0(u)} f_\theta'(u)\,du. \tag{6}$$

In the special case of deterministic exponential growth with rate λ, it follows from formula (5) that the random variable S has a translated exponential distribution, that is

$$\bar{F}_S(s) = \exp\left\{ -\frac{\alpha}{\lambda}(s-1) \right\}, \quad s \geqslant 1 \tag{7}$$

and therefore,

$$\mathbf{E}\{S\} = 1 + \frac{\lambda}{\alpha}.$$

Consider a new random variable V representing tumor volume at detection. To find the distribution of V we specify the law of tumor growth in volume units by the function $v(t) = ce^{\lambda t}$, where c is the volume of a single cell. Then it follows from (7) that

$$Pr\{V > v\} = \bar{F}_V(v) = \exp\left\{ -\frac{\alpha}{\lambda c}(v-c) \right\}, \quad v \geqslant c. \tag{8}$$

It is important to note that the distribution of tumor size (volume) at detection does not depend on the time of tumor onset.

The simplest way to generalize the above model is through allowing for random parameter θ. The random nature of the parameter θ can be interpreted as inter-individual variability of the law of tumor growth. Let $h(\theta)$, $\theta \geqslant 0$, be the prior distribution density of this parameter. From formula (5) we immediately obtain

$$\bar{F}_S(s) = \int_0^\infty e^{-\alpha\Phi_0(\psi_0(s))} h(\theta)\,d\theta. \tag{9}$$

It is interesting to mention that the law of tumor growth can be recovered from the distribution of tumor size at detection. To see this, denote by $p_S(s)$ to be the probability density function (p.d.f.)

of tumor size at detection. Differentiating formula (9) with respect to s, solving for the derivative of ψ_θ, and then integrating the resultant expression, we have

$$\psi_\theta(s) = \frac{1}{\alpha}\int_1^s \frac{p_S(u)}{\bar{F}_S(u)}\frac{\mathrm{d}u}{u}. \tag{10}$$

If one is interested in tumor volume rather than the number of tumor cells at detection, then formula (10) assumes the form

$$\psi_\theta(v) = \frac{c}{\alpha}\int_c^{cv} \frac{p_V(u)}{\bar{F}_V(u)}\frac{\mathrm{d}u}{u},$$

where $p_V(v) = c^{-1}p_S(v/c)$ is the p.d.f. of the tumor volume at detection and $\bar{F}_V(v) = \bar{F}_S(v/c)$ is the corresponding tail function.

If the growth of tumor volume is exponential with parameter λ and $\gamma = 1/\lambda$ follows a gamma distribution with shape parameter a and scale parameter b then by compounding (8) we find that the p.d.f. p_V of the tumor volume at detection follows a translated version of the generalized Pareto distribution

$$p_V(v) = \frac{a\beta}{a+1}\left[1+\frac{\beta}{a+1}(v-c)\right]^{-(a+1)}, \tag{11}$$

where $\beta = \alpha(a+1)/(bc)$.

Suppose we have a sample $v_1,\ldots,v_n$ of tumor volumes at diagnosis. The maximum likelihood estimates of a and b can be obtained by maximizing the log-likelihood

$$l = n\log a + n\log\tilde{\beta} - (a+1)\sum_{j=1}^{n}\log[1+\tilde{\beta}(v_j-c)] \tag{12}$$

expressed in terms of parameters a and $\tilde{\beta} := \beta/(a+1) = \alpha/(bc)$. The equations for computing the maximum likelihood estimates are as follows:

$$\left[\frac{1}{n}\sum_{j=1}^{n}\frac{1}{1+\tilde{\beta}(v_j-c)}\right]\left[1+\frac{1}{n}\sum_{j=1}^{n}\log[1+\tilde{\beta}(v_j-c)]\right] = 1, \quad a = \left[\frac{1}{n}\sum_{j=1}^{n}\log(1+\tilde{\beta}u_j)\right]^{-1}.$$

Observe that these equations always have a solution $\tilde{\beta} = 0$, $a = \infty$ which corresponds, as it follows from (11), to the translated exponential distribution. Examples show that other solutions may not exist or be multiple, see [16,17] for a more detailed discussion. In the case of multiple roots, the usual regularity conditions are no longer sufficient to guarantee consistency of the maximum likelihood estimator, even when it exists for all n [18]. A cure for this difficulty is to construct a $\sqrt{n}$-consistent estimator and then apply the first step of the Newton–Raphson iterative procedure. It can be shown that, under the commonly invoked regularity conditions, the resulting estimator sequence is consistent, asymptotically normal, and efficient [18]. The simplest way to find a $\sqrt{n}$-consistent estimator is by constructing the moment estimator (based on the first two moments) or applying the accumulation method [19] associated with a certain prescribed partition of the data range. It should be kept in mind, however, that the method of moments for the distribution (11) is feasible only for $a > 2$. Direct maximization of the log-likelihood function (12) represents an alternative way of finding the maximum likelihood estimators of the parameters

incorporated into the generalized Pareto distribution [20]. Since c is at least six orders of magnitude less than the smallest of the vs, one may disregard it in formula (12), if desired. A detailed study of maximum likelihood estimation for the distribution (11) in comparison to the method of moments is given in [21].

The way of estimation of the parameters a and b described above implies that the coefficient α is known. Alternatively, one can apply the same randomization procedure not to the parameter $\gamma = 1/\lambda$ but to the product $\alpha\gamma$, thinking of the latter as a gamma-distributed random variable with shape parameter a_1 and scale parameter b_1. The resultant distribution has the following density:

$$p_V(v) = \frac{a_1\beta_1}{a_1+1}\left[1+\frac{\beta_1}{a_1+1}(v-c)\right]^{-(a_1+1)}, \tag{13}$$

where $\beta_1 = (a+1)/(b_1 c)$.

To see whether the above randomized version of the distribution (8) really makes a difference when applied to epidemiological data, we analyzed measurements of tumor size at detection in 1120 patients with lung cancer (all clinical stages) identified through the Utah Cancer Registry. These measurements are provided by pathological records. The method of maximum likelihood with direct maximization of the log-likelihood was used to estimate the numerical parameters incorporated into models (8) and (13).

The two parametric estimates of the p.d.f. of $\log V$ are compared with the corresponding histogram in Fig. 1. The results of this analysis clearly indicate that the exponential distribution (8) is entirely inconsistent with the data under study. On the other hand, its randomized counterpart (13) provides a good fit to the data.

Another way of modifying the basic model is through a different choice of tumor growth kinetics. In particular, the Gompertz law of tumor growth is specified by the formula

$$f_{\delta_1,\delta_2}(t) = e^{\delta_1(1-e^{-\delta_2 t})}$$

with constant parameters $\delta_1, \delta_2 > 0$. The corresponding p.d.f. of tumor size at detection is given by

$$\begin{aligned} p_S(s) &= \frac{\alpha}{\delta_2(\delta_1 - \log s)} \exp\left\{ -\alpha \int_0^{\frac{1}{\delta_2}[\log\delta_1 - \log(\delta_1 - \log s)]} e^{\delta_1(1-e^{-\delta_2 u})}\, du \right\} \\ &= \frac{\alpha}{\delta_2(\delta_1 - \log s)} \exp\left\{ -\frac{\alpha}{\delta_2}\int_1^s \frac{dx}{\delta_1 - \log x} \right\} \quad \text{for } 1 \leqslant s < e^{\delta_1}. \end{aligned}$$

We evaluated the effect of tumor growth kinetics on the shape of the p.d.f. of tumor size by sample computations. The p.d.f. p_S for the Gompertz growth depends on two parameters α/δ_2 and δ_1. We selected the following numerical values of these parameters: $\alpha/\delta_2 = 1.4 \times 10^{-10}$ and $\delta_1 = 25$. To see how well can the p.d.f. p_S be approximated by the exponential density (derived from formula (7)) of the form

$$p_S^*(s) = \frac{\alpha}{\lambda}\exp\left\{-\frac{\alpha}{\lambda}(s-1)\right\}, \quad s \geqslant 1,$$

we used the transformation $Z = \log S$ and then minimized the Hellinger distance [22] between the two densities with respect to the parameter α/λ. The distance attained its minimum for

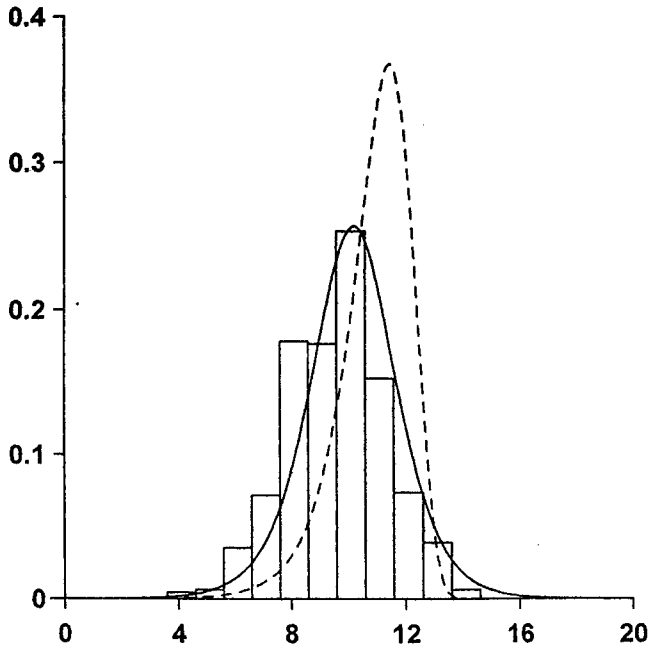


Fig. 1. The probability density function for $\log V$ estimated from data on the volume V of lung cancer at diagnosis. Dashed line represents the parametric estimate based on formula (8), solid line represents the estimate based on the more general model (13), step-wise curve is the histogram constructed from the sample values of $\log V$, where V is measured in mm^3.

$\alpha/\lambda = 7.67 \times 10^{-11}$, which value was used when computing the p.d.f. of Z in the case of exponential growth. Fig. 2 shows quite dissimilar shapes of the distribution of tumor size at detection for the two different kinetic curves of tumor growth.

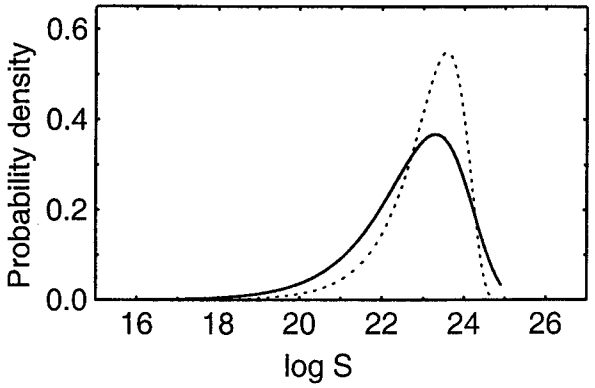


Fig. 2. The p.d.f. of $\log S$ for two different types of tumor growth. Solid line: exponential growth, dashed line: Gompertzian growth.

5. Age at detection and its distribution

Given the tumor growth begins at time $t = 0$, the survival function for the time to detection is specified by formula (4). In particular, for non-random exponential tumor growth with rate λ, we have

$$\bar{G}_W(w) = e^{-\frac{g}{\lambda}(e^{\lambda w}-1)}, \quad w \geqslant 0. \tag{14}$$

A randomized counterpart of $\bar{G}_W(w)$ may be written as

$$\bar{G}_W(w) = \int_0^\infty e^{-\alpha y(e^{w/y}-1)} h(y)\,dy, \quad w \geqslant 0, \tag{15}$$

where h represents the p.d.f. of the parameter $\gamma = 1/\lambda$. We used a gamma distribution for γ in our analysis of the data on lung cancer based on the distribution of tumor size at detection (Section 4). This analysis resulted in an estimate of the shape parameter of the function h that was very close to 1, thereby suggesting an exponential distribution, $h(y) = b\exp(-by)$, $y > 0$, to be used in the compounding procedure. Setting the coefficient α equal to 2.3×10^{-10} we obtained the estimate $b = 6.9$. The resultant distribution density p_W is plotted in Fig. 3. Shown in this figure are two other densities p_W computed for two different values of α. It is clear that even a two orders of magnitude difference in the value of α (given the same value of b) does not significantly change the shape of the function p_W.

In describing the natural history of cancer, the process of tumor development can be broken down into three stages. These stages are:

- formation of initiated cells;
- promotion of initiated cells resulting in appearance of the first malignant clonogenic cell;
- subsequent growth and progression of malignant tumor.

The duration of each stage of carcinogenesis is thought of as a random variable.

The above classification suggests that the event of malignant transformation occurs at some random time T representing the total duration of the first two stages (initiation and promotion) of carcinogenesis. In the case of sporadic carcinogenesis the r.v. T is measured from the date of birth of an individual. Let $\bar{G}_T$ be the survival function of the r.v. T, and let g_T stand for its density. There are mechanistically motivated models of carcinogenesis that yield an explicit expression of the survival function $\bar{G}_T(t)$ or the corresponding hazard function.

The most widely accepted two-stage model of carcinogenesis is commonly referred to as the Moolgavkar–Venzon–Knudson (MVK) model [23–25]. The standard form of this Markovian two-stage model involves four parameters $(\theta_0, \lambda_0, \mu_0, \eta_0)$ that refer to the rates of initiation of target stem cells (that is, formation of primary precancerous lesions), and rates of division, death (or differentiation), and malignant transformation of initiated and promoted cells. It was first pointed out by Heidenreich [26] and subsequently by Hanin and Yakovlev [27] and Heidenreich et al. [28] that these four parameters are not jointly identifiable from time-to-tumor data. In the case of constant parameters, all triples of their identifiable combinations have been described in [27]. In the latter case, the MVK model leads to the following explicit formula for the distribution of the total duration T of the first two stages, that is, of the time from the birth of an individual to the tumor onset [24,29–31]

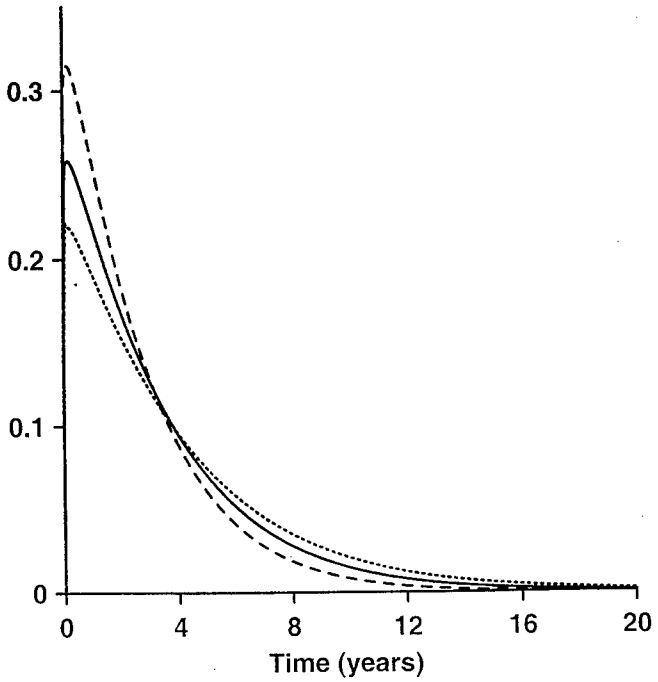


Fig. 3. The probability density functions p_W for the time at detection (progression stage duration) distribution at values of model parameters estimated from data on the volume of lung cancer at diagnosis and different values of the parameter α. Solid line: $\alpha = 2.3 \times 10^{-8}$, dashed line: $\alpha = 2.3 \times 10^{-10}$, dotted line: $\alpha = 2.3 \times 10^{-12}$.

$$\bar{G}_T(t) := Pr(T > t) = \left[\frac{(A+B)e^{At}}{B + Ae^{(A+B)t}}\right]^{\delta}, \quad t \geqslant 0. \tag{16}$$

Here A, B, $\delta > 0$ are the identifiable parameters of the model. In formula (16) we use the following identifiable parameterization:

$$\delta = \theta_0/\lambda_0, \quad A = \sqrt{(\lambda_0 + \mu_0 + \eta_0)^2 - 4\lambda_0\mu_0} + (\mu_0 + \eta_0 - \lambda_0),$$
$$B = \sqrt{(\lambda_0 + \mu_0 + \eta_0)^2 - 4\lambda_0\mu_0} - (\mu_0 + \eta_0 - \lambda_0).$$

Properties of the hazard function of the r.v. T distributed in accordance with the survival function (16) are described in [28].

Fig. 4 shows a typical p.d.f. g_T that corresponds to the survival function (16). In these computations, we used the values of A, B, and δ for lung cancer which were estimated by Luebeck et al. [31] from the control group identified through the Colorado Uranium Miners Cohort. Specifically, we set $A = 10^{-4}, B = 0.1821, \delta = 0.0364$ in these computations. Numerical algorithms are available for computing $\bar{G}_T$ in the case of piece-wise constant rates θ_0, λ_0, μ_0, η_0 [24,32,33].

Another model of carcinogenesis was proposed by Yakovlev and Polig [34]. The key feature of the Yakovlev–Polig (Y–P) model is that it allows for the process of cell death to compete with the process of tumor promotion. According to this model, the hazard function ϕ of the time T of tumor latency, which is related to the survival function by

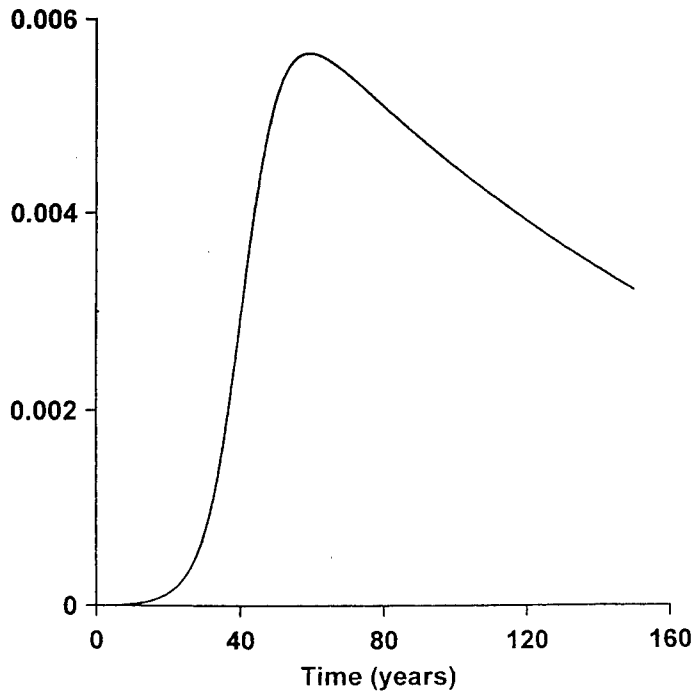


Fig. 4. The probability density function g_T for the time-to-onset distribution based on the MVK model of sporadic carcinogenesis. Computations were carried out at the following estimates of model parameters obtained from the Colorado Uranium Miners Cohort [32]: $\hat{A} = 10^{-4}, \hat{B} = 0.1821, \hat{\delta} = 0.0364$.

$$\bar{G}_T(t) = \exp\left\{ -\int_0^t \phi(u)\,\mathrm{d}u \right\}, \quad t \geqslant 0 \tag{17}$$

is of the form

$$\phi(u) = \theta_1 \exp\left\{ -\theta_2 \int_0^u l(x)\,\mathrm{d}x \right\} \int_0^u l(x)p(u-x)\,\mathrm{d}x, \quad u \geqslant 0, \tag{18}$$

where l is a given time-dependent rate of external exposure, p is the marginal (with respect to the joint distribution of the promotion time and the time to cell death for initiated cells) p.d.f. of the tumor promotion time, and $\theta_1 > 0$, $\theta_2 \geqslant 0$ are constants. The usual practice is to use a flexible parametric family of distributions for the function p, for example the two-parameter gamma distribution.

The hazard function ϕ has a maximum whenever $\theta_2 > 0$ and either of the functions l and p is bounded. In the case where $\int_0^\infty l(x)\,\mathrm{d}x$ is finite, this assertion was proven in [34]. If $\int_0^\infty l(x)\,\mathrm{d}x$ is infinite and p is bounded (almost everywhere) from above by a constant C, we have

$$\phi(t) \leqslant C\theta_1 \left(\int_0^t l(x)\,\mathrm{d}x \right) \exp\left\{ -\theta_2 \int_0^t l(x)\,\mathrm{d}x \right\};$$

hence $\phi(t) \to 0$ as $t \to \infty$. It is easy to see that ϕ displays the same behavior if the function l is bounded from above. Since $\phi(0) = 0$ and $\phi(t) > 0$ for $t > 0$, the function ϕ must have a maxi-

mum. However, the situation is not the same if we set θ_2 equal to zero. Suppose in addition that $l(x) = 1$ for almost all $x \geqslant 0$, which is a reasonable assumption when modeling spontaneous carcinogenesis. Then, it follows from (18) that

$$\phi(t) = \theta_1 \int_0^t p(x)\,dx,$$

that is, ϕ is a non-decreasing function of time.

Recently, Hanin and Boucher [35] found conditions under which the model given by (17) and (18) is identifiable from time-to-tumor observations. The model has been applied to various sets of experimental and epidemiological data to gain quantitative insight into the process of tumorigenesis induced by radiation [36–42] and chemical carcinogens [43–47]. The model also explains some peculiarities in the incidence of female colorectal cancer [48].

Both models can be represented [27] by the following general formula for the survival function $\bar{G}_T$:

$$\bar{G}_T(t) = \exp\left\{ -\theta_0 \int_0^t K(u)\,du \right\}, \tag{19}$$

where θ_0 is the rate of initiation, and K is the promotion time cumulative distribution function. More specifically, $K(u)$ defines the probability that a cell initiated at time 0 completes the promotion stage at or before time u. The distribution K is improper if a given two-stage model allows for cell death in the course of tumor promotion. Formula (19) extends to the case of time-dependent initiation rate $\theta(t)$ as follows [38]:

$$\bar{G}_T(t) = \exp\left\{ -\int_0^t \theta(t-u)K(u)\,du \right\}. \tag{20}$$

Mechanistic modeling of the process of tumor detection suggests a natural way of incorporating the progression stage into the existing mechanistic two-stage models of carcinogenesis. When considering some common human malignant tumors, it seems plausible to assume that the event of malignant transformation is rare and the process of tumor detection is triggered by the first arrival of a malignant cell. In other words, once the first malignant cell is generated, its subsequent development (progression) into an overt tumor is irreversible and the detection process begins, but those cell clones that may be generated at later times do not contribute to the process. In this case, the time of tumor latency (age at tumor detection) can be represented as $U = T + W$. Assuming stochastic independence between T and W we obtain the p.d.f. g_U of U as the convolution

$$g_U(t) = \int_0^t g_T(t-\tau)p_W(\tau)\,d\tau. \tag{21}$$

Shown in Fig. 5 is a plot of the density g_U computed in accordance with formula (21) at the parameter values used earlier for computing the densities p_W and g_T (Figs. 3 and 4). In is natural that variations in the parameter α are even less tangible than those seen in Fig. 3.

However, the model represented by convolution (21) does not work well in settings where multiple tumors are observed. In such settings (for one example, see [47]), the development of nonlethal tumors is irreversible and the total number of tumors generated over any time interval $(0, t]$ is recorded. All practically used two-stage stochastic models of carcinogenesis are essentially

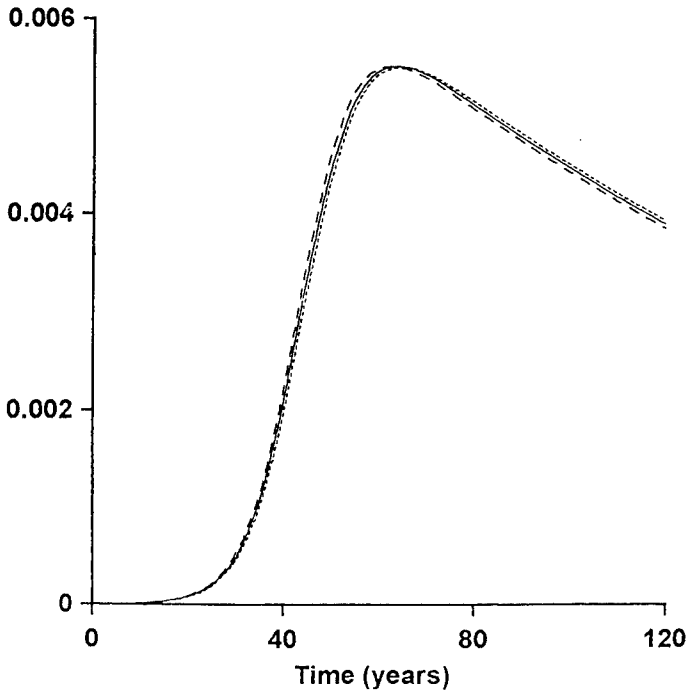


Fig. 5. The convolution of the p.d.f.s g_T (see Fig. 2) and p_W (Fig. 3) for lung cancer as a function of the parameter α. Solid line: $\alpha = 2.3 \times 10^{-8}$, dashed line: $\alpha = 2.3 \times 10^{-10}$, dotted line: $\alpha = 2.3 \times 10^{-12}$.

based on the simplifying assumption that the process of initiation can be modeled as a Poisson process with intensity $\theta(t)$. If initiated cells are promoted independently of one another, the number of cells initiated and subsequently promoted by time t is a Poisson process with intensity

$$\Lambda(t) = \int_0^t \theta(t-u)K(u)\,\mathrm{d}u,$$

see [27,34]. Introduce the convolution representing the cumulative distribution function of the total duration of promotion and progression stages:

$$[K * G_W](u) = \int_0^u K(u-w)\,\mathrm{d}G_W(w),$$

where G_W is the cumulative distribution function of the progression stage duration. Here we assume again that the promotion and progression stages are independent. It follows from the assumption on the Poisson process of initiation events that the survival function of the time to the first tumor is given by

$$\bar{G}(t) = \exp\left\{ -\int_0^t \theta(t-u)[K * G_W](u)\,\mathrm{d}u \right\},$$

see [27].

The latter formula has in fact much broader implications since it describes the situation where multiple events of tumor detection can be interpreted as independent 'competing risks'. Then it follows that the time to the first event of detection is given by a minimum of a random (Poisson distributed) number of independent and identically distributed random variables representing individual detection times of multiple tumors (or cell clones). Therefore, it makes sense to explore this model as an alternative to model (21) in a wider spectrum of applications, including lethal tumors.

6. Identifiability of the distribution of the time to detection

Let us consider now the following problem. Suppose that the distribution of the time to onset T either belongs to the gamma family or is specified according to the MVK model by formula (16). In both cases, parameters of the distribution of the r.v. T are identifiable from its observed distribution [27,35]. The same is true for the Y–P model under some mild conditions formulated in [35].

It immediately follows from formula (4) that the p.d.f. of the time to detection, p_W, is given by

$$p_W(w) = \alpha f(w) \exp\left\{ -\alpha \int_0^w f(s)\,\mathrm{d}s \right\}, \quad w \geqslant 0. \tag{22}$$

Assuming that tumor grows exponentially with a fixed rate λ, one can find on the basis of formula (22) that the parameters α and λ are identifiable within this model given the times of spontaneous tumor detection. A natural question is whether the entire set of parameters involved in the distribution of the age at spontaneous detection is identifiable. We limit our consideration to the model (21) representing the time of tumor latency U as the sum of the two independent random variables T and W. This leads to the following problem.

Problem. Let $\mathscr{P}$ and $\mathscr{G}$ be two families of probability distributions on $[0,\infty)$. Is it true that the family of convolutions $P * G$, where $P \in \mathscr{P}$ and $G \in \mathscr{G}$, is identifiable? In other words, does

$$P_1 * G_1 = P_2 * G_2,$$

where $P_1, P_2 \in \mathscr{P}$ and $G_1, G_2 \in \mathscr{G}$, imply that $P_1 = P_2$ and $G_1 = G_2$?

Taking both families to be the set of degenerate distributions $\delta_a, a \geqslant 0$, and observing that $\delta_a * \delta_b = \delta_{a+b}$, we conclude that in general the answer to this question is negative. Yet another counterexample is given by gamma distributions $\Gamma(a_1, b)$ and $\Gamma(a_2, b)$ with the convolution being equal to $\Gamma(a_1 + a_2, b)$. To formulate a theorem providing sufficient conditions for the positive solution of the problem we need the following definition, see [35].

Definition 2. A family of absolutely continuous distributions on $[0,\infty)$ is called *graduated* if for every two distinct p.d.f.'s q_1 and q_2 from this family and for every $\varepsilon > 0$ there exists a number $M > 0$ such that either $q_1(x) \leqslant \varepsilon q_2(x)$ for all $x > M$ or $q_2(x) \leqslant \varepsilon q_1(x)$ for all $x > M$.

In other words, a family is graduated if the limit at infinity of the ratio of any two distinct p.d.f.'s from the family is either 0 or infinity. It is easy to see that the gamma family is graduated. The same is true for the family

$$g_{\alpha,\lambda}(x) = \alpha e^{\lambda x} e^{-\frac{\alpha}{\lambda}(e^{\lambda x}-1)}, \quad x \geqslant 0; \quad \alpha, \quad \lambda > 0, \tag{23}$$

of p.d.f.'s of the form (22) with $f(w) = e^{\lambda w}$ corresponding to the survival functions (14).

Remark 1. Observe that family (16) is not graduated, because any p.d.f. from this family with parameters A, B, δ behaves at infinity like $C\exp\{-b\delta t\}$, where C is a positive constant depending on A, B and δ, so that the ratio of p.d.f.'s for two distinct distributions with parameters A_1, B_1, δ_1 and A_2, B_2, δ_2 satisfying $B_1\delta_1 = B_2\delta_2$ tends at infinity to a constant different from 0 and infinity. The same is true for any family of distributions that can be represented in the form of formula (19) if the distribution K has finite first moment.

Theorem 1. *Let $\mathscr{P}$ and $\mathscr{G}$ be two families of absolutely continuous probability distributions on $[0,\infty)$ with p.d.f.'s $p \in \mathscr{P}$ and $g \in \mathscr{G}$. Suppose that*

1. *family $\mathscr{P}$ is graduated;*
2. *for every $p \in \mathscr{P}$, there is $M = M(p) > 0$ such that $p(t) > 0$ for all $t > M$;*
3. *for every $p \in \mathscr{P}$ and for each $s > 0$, there exists a finite limit*

$$h_p(s) := \lim_{t\to\infty} \frac{p(t-s)}{p(t)};$$

4. *for each $g \in \mathscr{G}$,*

$$\lim_{t\to\infty} \int_0^t \frac{p(t-s)}{p(t)} g(s)\,\mathrm{d}s = \int_0^\infty h_p(s)g(s)\,\mathrm{d}s;$$

5. *for all $p \in \mathscr{P}$ and $g \in \mathscr{G}$,*

$$0 < \int_0^\infty h_p(s)g(s)\,\mathrm{d}s < \infty.$$

Then the family of convolutions $P * G$, where $P \in \mathscr{P}$ and $G \in \mathscr{G}$, is identifiable.

Proof. Suppose that for some distributions $P_1, P_2 \in \mathscr{P}$ and $G_1, G_2 \in \mathscr{G}$ we have $P_1 * G_1 = P_2 * G_2$. Then

$$\int_0^t p_1(t-s)g_1(s)\,\mathrm{d}s = \int_0^t p_2(t-s)g_2(s)\,\mathrm{d}s, \quad t \geqslant 0.$$

Assuming that $t > M := \max\{M(p_1), M(p_2)\}$ we rewrite this equation in the form

$$p_1(t)\int_0^t \frac{p_1(t-s)}{p_1(t)} g_1(s)\,\mathrm{d}s = p_2(t)\int_0^t \frac{p_2(t-s)}{p_2(t)} g_2(s)\,\mathrm{d}s, \quad t > M.$$

Passage to limit as $t \to \infty$ in this equation with conditions (3)–(5) of the theorem taken into account yields

$$\lim_{t\to\infty} \frac{p_1(t)}{p_2(t)} = \frac{\int_0^\infty h_{p_2}(s)g_2(s)\,\mathrm{d}s}{\int_0^\infty h_{p_1}(s)g_1(s)\,\mathrm{d}s}.$$

Invoking conditions (1) and (5) we conclude that $p_1 = p_2$. Denote this function by p.

For any p.d.f. φ of a non-negative random variable, let $\hat{\varphi}$ be its Laplace transform defined by

$$\hat{\varphi}(z) := \int_0^\infty e^{-zt}\varphi(t)\,dt, \quad Re\ z \geqslant 0.$$

Then

$$\hat{p}(z)\hat{g}_1(z) = \hat{p}(z)\hat{g}_2(z), \quad Re\ z \geqslant 0. \tag{24}$$

Since $\hat{p}$ is a non-zero analytic function in $\{Re\ z > 0\}$, the set $Q := \{t \in (0,\infty)\colon \hat{p}(t) \neq 0\}$ is open and non-empty. From (24) we conclude that $\hat{g}_1 = \hat{g}_2$ on Q and hence, by the uniqueness theorem for analytic functions, $\hat{g}_1(z) = \hat{g}_2(z)$ for all z with $Re\ z \geqslant 0$. Therefore, by the uniqueness theorem for the Laplace transform, $g_1 = g_2$. Thus, $P_1 = P_2$ and $G_1 = G_2$. Theorem 1 is proved.

Remark 2. For concrete families $\mathscr{P}$ and $\mathscr{G}$, condition (4) usually follows from standard theorems about passage to limit in the Lebesgue integral. In particular, Theorem 1 can be applied in the case when $\mathscr{P}$ is the family of gamma distributions $\Gamma(a,b)$ with shape parameter $a \geqslant 1$ and $\mathscr{G}$ is the family (23). First, conditions (1) and (2) of Theorem 1 are obviously satisfied. Next, for $p \in \Gamma(a,b)$ with $a \geqslant 1$ we have for all $s \geqslant 0$

$$\frac{p(t-s)}{p(t)} = \left(\frac{t-s}{t}\right)^{a-1} e^{bs} \to e^{bs} \quad \text{as} \quad t \to \infty.$$

Hence, condition (3) is met with $h_p(s) = \exp\{bs\}$. Further, assuming that $a \geqslant 1$ we obtain, for any p.d.f. g from the family (23),

$$\int_0^t \frac{p(t-s)}{p(t)} g(s)\,ds = \int_0^\infty \frac{p(t-s)}{p(t)} \chi_{[0,t]}(s) g(s)\,ds \to \int_0^\infty e^{bs} g(s)\,ds < \infty$$

by the Lebesgue theorem on dominated convergence, which is condition (4). Finally, condition (5) also holds. This leads us to a conclusion that if promotion time T has a gamma distribution $\Gamma(a,b)$ with $a \geqslant 1$ and tumor growth is exponential with rate λ, then parameters a, b, α and λ are jointly identifiable from the observed distribution of the age at spontaneous detection.

Remark 3. It is easy to check that the convolution of the density corresponding to the MVK model given by (16) and the p.d.f. specified by (23) does not satisfy the sufficient conditions formulated in Theorem 1. The same is true for this convolution if the MVK model is replaced with the Y–P model. This does not mean, however, that these convolutions are non-identifiable, but more powerful analytical results are necessary to clarify their properties associated with the notion of identifiability.

7. Model stability

Suppose the law of tumor growth is described by an increasing function f. A natural question to ask is how sensitive is the distribution of the r.v. W given by (4) to the change of the law of tumor growth. Solving this problem presupposes the choice of two metrics that measure distances between the distributions of W and functions f, and establishing a relation between these metrics.

For two survival functions $\bar{F}_1, \bar{F}_2$ of non-negative random variables, denote

$$\rho_\infty(\bar{F}_1,\bar{F}_2) := \sup\{|\bar{F}_1(t)-\bar{F}_2(t)| : t \geqslant 0\}. \tag{25}$$

Obviously, this formula defines a probability metric. Next, let $\mathscr{F}$ be the set of all non-negative increasing functions on $[0,\infty)$. Given $\alpha > 0$, define, for any functions $f_1, f_2 \in \mathscr{F}$,

$$d_\alpha(f_1,f_2) := \inf_{T>0} \max\left\{\alpha\int_0^T |f_1(t)-f_2(t)|\,dt, \right.$$
$$\left. \exp\left(-\alpha\int_0^T f_1(t)\,dt\right), \quad \exp\left(-\alpha\int_0^T f_2(t)\,dt\right)\right\}. \tag{26}$$

It is readily checked that d_α is a metric on $\mathscr{F}$. Also, it is easy to verify that for all $f_1, f_2 \in \mathscr{F}, f_1 \neq f_2$, there exists a unique number $T_0 = T_0(f_1,f_2)$ such that

$$d_\alpha(f_1,f_2) = \alpha\int_0^{T_0} |f_1(t)-f_2(t)|\,dt$$
$$= \max\left\{\exp\left(-\alpha\int_0^{T_0} f_1(t)dt\right), \quad \exp\left(-\alpha\int_0^{T_0} f_2(t)\,dt\right)\right\}.$$

Theorem 2. *Let $f_1, f_2 \in \mathscr{F}$ be two laws of tumor growth, and let $\bar{F}_1, \bar{F}_2$ be the corresponding survival functions defined in* (4). *Then*

$$\rho_\infty(\bar{F}_1,\bar{F}_2) \leqslant d_\alpha(f_1,f_2).$$

Proof. Fix $T > 0$. Denote

$$d_\alpha(f_1,f_2;T) := \max\left\{\alpha\int_0^T |f_1(t)-f_2(t)|\,dt,\ \exp\left(-\alpha\int_0^T f_1(t)\,dt\right),\ \exp\left(-\alpha\int_0^T f_2(t)\,dt\right)\right\},$$

so that in view of (26)

$$d_\alpha(f_1,f_2) = \inf_{T>0} d_\alpha(f_1,f_2;T). \tag{27}$$

Observe that according to the mean value theorem for all $x,\ y \geqslant 0$,

$$|e^{-\alpha x} - e^{-\alpha y}| \leqslant \alpha|x-y|.$$

Therefore, for $0 \leqslant t \leqslant T$, we have

$$|\bar{F}_1(t)-\bar{F}_2(t)| = \left|\exp\left(-\alpha\int_0^t f_1(t)\,dt\right) - \exp\left(-\alpha\int_0^t f_2(t)\,dt\right)\right|$$
$$\leqslant \alpha\int_0^t |f_1(s)\,ds - \int_0^t f_2(s)\,ds| \leqslant \alpha\int_0^T |f_1(t)-f_1(t)|\,dt$$
$$\leqslant d_\alpha(f_1,f_2;T). \tag{28}$$

Next, for $t > T$,

$$|\bar{F}_1(t) - \bar{F}_2(t)| \leqslant \max\left\{\exp\left(-\alpha\int_0^t f_1(u)\,du\right), \exp\left(-\alpha\int_0^t f_2(u)\,du\right)\right\}$$
$$\leqslant \max\left\{\exp\left(-\alpha\int_0^T f_1(u)\,du\right), \exp\left(-\alpha\int_0^T f_2(u)\,du\right)\right\}$$
$$\leqslant d_\alpha(f_1, f_2; T). \tag{29}$$

Combining (28) and (29) we find that, for all $t \geqslant 0$,

$$|\bar{F}_1(t) - \bar{F}_2(t)| \leqslant d_\alpha(f_1, f_2; T), \quad T > 0.$$

Therefore, $\rho_\infty(\bar{F}_1, \bar{F}_2) \leq d_\alpha(f_1, f_2; T)$ for all $T > 0$, which in view of (27) concludes the proof of Theorem 2.

8. Joint distribution of age and tumor size at detection and its randomized form

Recall that T is the age at tumor onset, W is the time of spontaneous tumor detection measured from the onset of disease, and S is the tumor size at spontaneous detection. Then $S = f(W)$, where $f : [0, \infty) \to [1, \infty)$ is a deterministic function describing the law of tumor growth. It is assumed that

1. random variables T and W are absolutely continuous and independent;
2. function f is differentiable and $f' > 0$;
3. the rate of spontaneous tumor detection is proportional to the current tumor size with coefficient $\alpha > 0$.

It follows from Assumption 3 that the p.d.f. of the random variable W is described by formula (22).

We observe sample values of the random vector $Y := (T + W, S)$ whose components are interpreted as age and tumor size at spontaneous detection, respectively. We look at Y as a transformation of the random vector $X := (T, W)$, $Y = \varphi(X)$, where $\varphi(t, w) = (t + w, f(w))$, $t, w \geqslant 0$. Observe that components of X are independent random variables. The inverse function $\psi = \varphi^{-1} : A \to \mathbb{R}^2_+$, where $A := \{(u, v) \in \mathbb{R}^2_+ : 1 \leqslant v \leqslant f(u)\}$, is given by $\psi(u, v) = (u - g(v), g(v))$, with $g := f^{-1}$. Note that the Jacobian of ψ is g'. Then for the p.d.f. of Y we have assuming that $(u, v) \in A$

$$p_Y(u, v) = p_X(\psi(u, v))g'(v)$$
$$= p_T(u - g(v))p_W(g(v))g'(v)$$
$$= p_T(u - g(v))p_S(v).$$

In the particular case of exponential tumor growth with rate $\lambda > 0$ ($f(w) = e^{\lambda w}$) we obtain using formula (7)

$$p_Y(u, v) = \frac{\alpha}{\lambda} e^{-\frac{\alpha}{\lambda}(v-1)} p_T\left(u - \frac{\ln v}{\lambda}\right), \quad u \geqslant 0, \ 1 \leqslant v \leqslant e^{\lambda u}. \tag{30}$$

Thus, the distribution of random vector Y is absolutely continuous but the support of Y depends on the unknown parameter λ. As far as the asymptotic likelihood inference is concerned, the usual

regularity conditions are not met for the distribution p_Y. However, experience with similar parametric settings suggests that the estimation efficiency for the parameter λ may be expected to be even higher than in the regular case although asymptotic normality may fail.

Suppose the distribution of the time of tumor latency T is known. Let $\{(u_i, v_i)\colon 1 \leqslant i \leqslant n\}$ be sample data on age and tumor size at detection. The structure of the joint distribution (30) suggests the following maximum likelihood procedure for estimation of the parameters α and λ:

(1) Denote $\theta = \alpha/\lambda$ in formula (30), and find the maximum likelihood estimate, $\hat{\theta}$, of the parameter θ using only the tumor size data $\{v_i : 1 \leqslant i \leqslant n\}$. It follows (see below) that the sample $\{v_i\}$ is drawn from an exponential distribution with parameter θ, and consequently

$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} v_i - 1}.$$

(2) Maximize the function

$$L(\lambda) = \prod_{i=1}^{n} p_T\left(u_i - \frac{\ln v_i}{\lambda}\right), \quad u_i > 0, \ v_i \geqslant 1,$$

or its logarithm, to find the estimate of λ denoted by $\hat{\lambda}$.

(3) The maximum likelihood estimate of α is given by $\hat{\alpha} = \hat{\theta}\hat{\lambda}$.

The above procedure does the same job as maximizing the likelihood function based on the joint distribution (30). To show this, let the joint density of the random variables U and V be of the form

$$p(u, v; \lambda, \theta) = g(v; \theta) f\left(u - \frac{\varphi(v)}{\lambda}\right)$$

with $u > 0$, $v \geqslant 1$, $\lambda > 0$, and $\varphi : [1, \infty) \to (0, \infty)$. It is assumed that $g(x) > 0$ for $x \geqslant 1$, $f(t) > 0$ for $t > 0$, and $f(t) = 0$ for $t \leqslant 0$. Suppose that there exists a unique maximizer $(\hat{\lambda}, \hat{\theta})$ for the likelihood function

$$L(\lambda, \theta) = \prod_{i=1}^{n} p(u_i, v_i; \lambda, \theta).$$

It is clear that $\hat{\lambda}$ and $\hat{\theta}$ are unique maximizers for the functions

$$L_1(\lambda) = \prod_{i=1}^{n} f\left(u_i - \frac{\varphi(v_i)}{\lambda}\right) \quad \text{and} \quad L_2(\theta) = \prod_{i=1}^{n} g(v_i; \theta),$$

respectively. Conversely, if $\hat{\lambda} > 0$ and $\hat{\theta}$ are unique maximizers for these functions, the pair $(\hat{\lambda}, \hat{\theta})$ is a unique maximizer for the likelihood function $L(\lambda, \theta)$. Finally, observe that $g(v; \theta)$ is the marginal density of the random variable V. Indeed, we have

$$\int_0^\infty f\left(u - \frac{\varphi(v)}{\lambda}\right) g(v; \theta)\, \mathrm{d}u = g(v; \theta) \int_0^\infty f(t)\, \mathrm{d}t = g(v; \theta).$$

The performance of the above-described estimation procedure was studied by computer simulations. A total of 50 pseudo-random samples of (u_i, v_i) were generated from the joint distribution (30); each sample contained $n = 100$ realizations of the random vector (U, V). We used the

composition method [49] to simulate samples of pairs (u_i, v_i). In accordance with this method, we first draw v_i from the marginal distribution of the random variable V, and then generate u_i from the distribution of U conditional on $V = v_i$. The p.d.f. p_T was specified by the MVK model with the survival function given by formula (16). We used the following values of model parameters: $\alpha = 2.3 \times 10^{-10}, \lambda = 6.9, A = 10^{-4}, B = 0.1821, \delta = 0.0364$. These values are suggested by the analysis of lung cancer data (Sections 4 and 5).

8.1. Simulation Experiment 1

In this experiment, we kept the parameters A, B, and δ at their true values and applied the estimation procedure to simulated data in order to obtain estimates of the parameters λ and α. In this case, the likelihood function can be maximized by a univariate search for λ with a fixed value of θ. The estimates of λ and α which resulted from each of the 50 samples were summarized by calculating their sample means $\bar{\lambda}$ and $\bar{\alpha}$, as well as the corresponding standard errors (of the sample mean) denoted by $\sigma_{\bar{\lambda}}$ and $\sigma_{\bar{\alpha}}$, respectively. We obtained the following numerical values: $\bar{\lambda} = 7.45$, $\sigma_{\bar{\lambda}} = 0.9$, $\bar{\alpha} = 2.53 \times 10^{-10}$, $\sigma_{\bar{\alpha}} = 0.34 \times 10^{-10}$. These results testify that, given the parameters A, B and δ are known, the estimation procedure performs well when applied to finite but sufficiently large samples.

8.2. Simulation Experiment 2

Proceeding from the same true parameter values, the estimation procedure was applied to simulated data to obtain estimates of all the parameters incorporated into the model. We used algorithm FLEXI [50] to maximize the corresponding likelihood function. Since there were three additional parameters to be estimated from simulated data, the size of each sample was increased up to 1000. The results were summarized in just the same way as in Experiment 1 to give: $\bar{\lambda} = 9.4$, $\sigma_{\bar{\lambda}} = 0.9$, $\bar{\alpha} = 3.1 \times 10^{-10}$, $\sigma_{\bar{\alpha}} = 3.1 \times 10^{-11}$, $\bar{A} = 9.5 \times 10^{-4}$, $\sigma_{\bar{A}} = 3.6 \times 10^{-4}$, $\bar{B} = 0.1407$, $\sigma_{\bar{B}} = 0.0599$, $\bar{\delta} = 0.0507$, $\sigma_{\bar{\delta}} = 0.006$. The 50 estimates of the p.d.f. $p_T(t)$ with the above estimates substituted for its parameters were averaged for each value of t. The resultant average (arithmetic mean) agrees quite closely with the true p.d.f. p_T as seen in Fig. 6.

8.3. Simulation Experiment 3

The estimation procedure was applied to a single sample of size 50 000 generated from the joint distribution (30). The estimated parameter values were: $\hat{\lambda} = 6.7$, $\hat{\alpha} = 2.24 \times 10^{-10}$, $\hat{A} = 5.1 \times 10^{-4}$, $\hat{B} = 0.1390$, $\hat{\delta} = 0.0475$. The true p.d.f. p_T is compared with its parametric estimate in Fig. 7.

The above simulation experiments show that estimation of the whole set of model parameters is feasible given the model is adequate for the processes under study, but obtaining unbiased estimates would require large sample sizes.

Suppose now that the process of tumor growth is described by the exponential law $f(w) = e^{\lambda w}$, $w \geqslant 0$, with a *random* growth rate λ. We also assume that the random parameter $\theta := \alpha/\lambda$ is gamma distributed with parameters a and b. Compounding (30) with respect to the gamma distribution of the parameter θ we find the p.d.f. of the resulting randomized distribution of the vector Y

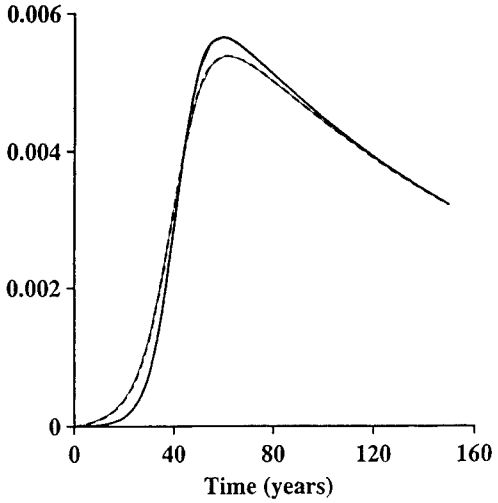


Fig. 6. Simulation Experiment 2: Comparison of the true p.d.f. p_T specified by the MVK model (solid line) with the corresponding estimate (dashed line) obtained by averaging over 50 parametric estimates of the p.d.f. p_T.

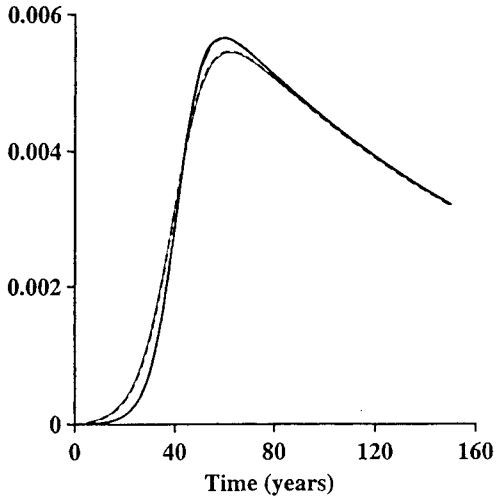


Fig. 7. Simulation Experiment 3: Comparison of the true p.d.f. p_T specified by the MVK model (solid line) with the corresponding estimate (dashed line) obtained from a single sample of 50 000 pairs of observations drawn from the joint distribution (30).

$$p(u,v) = \frac{b^a}{\Gamma(a)} \int_0^{\alpha u/\ln v} t^a \mathrm{e}^{-(b+v-1)t} p_T\left(u - \frac{\ln v}{\alpha} t\right) \mathrm{d}t, \quad u \geqslant 0, \ v \geqslant 1.$$

Setting $s := u - (\ln v/\alpha)t$ we rewrite the last formula in an equivalent form

$$p(u,v) = \frac{b^a}{\Gamma(a)} \left(\frac{\alpha}{\ln v}\right)^{a+1} \int_0^u (u-s)^a \exp\left\{-\frac{\alpha}{\ln v}(b+v-1)(u-s)\right\} p_T(s)\,\mathrm{d}s \tag{31}$$

for $u \geqslant 0$, $v \geqslant 1$. Alternatively, we may assume that it is the parameter $1/\lambda$ that is gamma distributed with parameters a and b. Should this be the case, we would have

$$p(u,v) = \frac{\alpha b^a}{\Gamma(a)} \int_0^{u/\ln v} t^a \exp\{-[b+\alpha(v-1)]t\} p_T(u - t\ln v)\,\mathrm{d}t$$
$$= \frac{\alpha b^a}{(\ln v)^{a+1}\Gamma(a)} \int_0^u (u-s)^a \exp\left\{-\frac{b+\alpha(v-1)}{\ln v}(u-s)\right\} p_T(s)\,\mathrm{d}s \tag{32}$$

for $u \geqslant 0$, $v \geqslant 1$.

Once the density p_T of the age at tumor onset T is specified within a certain parametric family (e.g. using a gamma distribution or the MVK model), Eqs. (31) or (32) allow us to compute p.d.f. of the joint distribution of age and tumor size at detection. Observe that in this randomized version the support $[0,\infty) \times [1,\infty)$ of the distribution of random vector Y is parameter-free. The maximum likelihood parametric inference based on the joint p.d.f. $p(u,v)$ accommodates censored observations under the usual censorship model [51].

9. A threshold model of tumor detection

An alternative approach to stochastic modeling of spontaneous detection was proposed and extensively discussed in [38,52,53]. The main postulate of the model developed in these works is that a tumor becomes detectable when its size attains some threshold value, N, which is treated as a random variable. The authors used a linear pure birth process with random absorbing upper barrier N to model the dynamics of tumor growth. Under this model the progression time cumulative distribution function, given the threshold level N, is

$$F(t|N) = (1 - \mathrm{e}^{-\lambda t})^{N-1}, \tag{33}$$

where λ is the birth rate. Formula (33) implies that tumor growth starts from a single malignant cell at time $t = 0$.

As mentioned in Section 4, it is practical to represent the critical number of tumor cells as $N = mV$, where V is the volume of a tumor, and $m = 1/c$ is the concentration of tumor cells per unit volume. The constant m is non-random and its values are typically large. Thus the conditional progression time c.d.f., given the threshold volume $V = v$, is

$$F(t|v) = (1 - \mathrm{e}^{-\lambda t})^{mv-1}. \tag{34}$$

Let $f(t|v)$ stand for the p.d.f. of $F(t|v)$.

Let $\bar{G}(t)$ be the survival function of the time it takes for the initiation and promotion processes to result in the event of neoplastic transformation. Assuming that the initiation rate is constant, i.e. $\theta(t) = \theta$, and the stages of promotion and progression are mutually independent, the authors of [52] used the convolution

$$g(t|v) = \theta \int_0^t K(t-u) \exp\left\{-\theta \int_0^{t-u} K(x)\,\mathrm{d}x\right\} f(u|v)\,\mathrm{d}u \tag{35}$$

to represent the conditional p.d.f., $g(t|v)$, of the time of tumor latency measured from the date of birth of an individual (see formulas (19) and (21)).

Introducing a prior distribution, $P(v)$, of the random variable V, we represent the p.d.f. of the time (age) of tumor detection as

$$g(t) = \int_0^\infty g(t|v)p(v)\,dv, \tag{36}$$

where $p(v)$ is the density of $P(v)$, and the distribution $P(v)$ is assumed to have finite first moment. One is primarily interested in the conditional p.d.f. of tumor volume at detection (given a tumor is detected at time t), hereafter denoted by $w(v|t)$. By virtue of Bayes' formula we have

$$w(v|t) = \frac{g(t|v)p(v)}{\int_0^\infty g(t|u)p(u)\,du} = \frac{g(t|v)p(v)}{g(t)}, \tag{37}$$

where $g(t|v)$ and $g(t)$ are given by (35) and (36), respectively.

This model yields an interesting asymptotic result showing that the conditional p.d.f. $w(v|t)$ assumes a very simple form when t tends to infinity. This limiting form does not involve the promotion time distribution K and it also has some distinct advantages as far as estimation problems are concerned [53]. It follows from (34) and (35) that

$$g(t|v) = \lambda\theta mv \int_0^t e^{-\lambda(t-s)}(1 - e^{-\lambda(t-s)})^{mv-1}K(s)\exp\left\{-\theta\int_0^s K(x)\,dx\right\}ds = \lambda\theta mv\psi(t).$$

Proven in [53] is the following theorem.

Theorem 3. *The following assertions hold for the limiting behavior of the function $\psi(t)$ as $t \to \infty$:*

1. *If $\lambda < \theta$, then*

$$\psi(t) \sim Ie^{-\lambda t},$$

where

$$I = \int_0^\infty \exp\left\{\lambda s - \theta\int_0^s K(x)\,dx\right\}K(s)\,ds.$$

2. *If $\lambda = \theta$ and $\int_0^\infty [1 - K(s)]\,ds < \infty$, then*

$$\psi(t) \sim t\exp\left\{-\lambda\int_0^t K(x)\,dx\right\}.$$

3. *If $\lambda > \theta$, then*

$$\psi(t) \sim J\exp\left\{-\theta\int_0^t K(x)\,dx\right\}$$

with

$$J = \frac{1}{\lambda}\int_0^1 y^{mv-1}(1-y)^{-\theta/\lambda}\,dy = \frac{1}{\lambda}B(mv, 1-\theta/\lambda),$$

where $B(x,y)$ is the beta function.

Corollary 1. *If $mv \to \infty$, it follows from Theorem 2 that the limiting conditional p.d.f. of tumor size at detection is of the form*

$$\lim_{t\to\infty} w(v|t) = \frac{v^{\mu}p(v)}{\int_0^{\infty} u^{\mu}p(u)\,du}, \quad \mu = \min\left\{1, \frac{\theta}{\lambda}\right\}.$$

A special case ($\mu = 1$) of this distribution is known as the length-biased sampling distribution in the theory of stationary point processes [54]. Analysis of data on breast cancer [53] revealed that the limiting distribution is quite adequate starting with the age of 50.

In studying stability of the posterior p.d.f. of tumor volume at detection under perturbation in the prior p.d.f. p [53], the following metric:

$$\rho_{\mu}(\tilde{f}, f) := \int_0^{\infty} u^{\mu}|\tilde{f}(u) - f(u)|\,du$$

is a natural choice. The corresponding theorem and some other stability results are described at length in [53]. An extended discussion of the above-described model in the context of parametric analysis of clinical data on breast cancer is presented in [55].

10. Modeling metastatic process

Metastatic progression of cancer can be modeled proceeding from the assumption that shedding a metastasis is a quantal response event. From the practical point of view, it seems important to derive a formula that gives the probability that at the time of detection, the primary tumor has not yet metastasized. This problem is tractable if we resort to a deterministic description of tumor growth. In doing so, we still can incorporate an additional randomness into the model by allowing for random values of its numerical parameters.

To derive a useful formula for the probability of the event of interest suppose that, in the course of growth, the tumor produces metastases with an intensity which is proportional (with coefficient ζ) to its current size $N(t)$. We assume that $N(t)$ is a non-random function of time. Specifically, we assume that the process of shedding metastases by the primary tumor is Poisson with intensity $\zeta N(t)$. Each metastasis is assumed to develop independently of the subsequent growth of the primary tumor and of other metastases. Any given metastasis grows deterministically and the rate of its detection is proportional (with coefficient α_2) to the current metastasis volume. In other words, given the time of metastasis origination τ, the rate of its detection at time $t > \tau$ equals $\alpha_2 M(t-\tau)$, where M is the size of the metastasis. The primary tumor is detected at a random time W_1 measured from the time of tumor onset $t = 0$. The rate of the primary tumor detection at time t is equal to $\alpha_1 N(t)$. Let W_2 be the time (also measured from $t = 0$) to the detection of the first metastasis.

Yorke et al. [56] considered a model of metastatic spread under a deterministic threshold mechanism of tumor detection. The authors used similar assumptions on growth characteristics of primary and metastatic tumors, but they did not evaluate the probability $Pr\{W_1 < W_2\}$ in their computations.

Note that

$$\pi(t) = e^{-\alpha_2 \int_0^t M(u)\,du} \tag{38}$$

is the probability that a metastasis of age t remains undetected. The probability that the primary tumor produces k metastases during the time t is

$$p_k(t) = \frac{1}{k!}\left(\zeta \int_0^t N(u)\,du\right)^k \exp\left\{-\zeta \int_0^t N(u)\,du\right\}, \quad k \geqslant 0.$$

Let τ_i be the time of the ith metastasis formation given that the number of metastases produced by time t is equal to k and $i \leqslant k$. Now we use the well-known fact that $\tau_1, \ldots, \tau_k$ are independent and identically distributed random variables, having the common density

$$\rho(u) = \begin{cases} \frac{N(u)}{\int_0^t N(z)\,dz}, & 0 \leqslant u \leqslant t, \\ 0, & u > t, \end{cases} \tag{39}$$

see [57]. Then

$$\begin{aligned} \Pi &= Pr\{W_1 < W_2\} \\ &= \int_0^\infty \alpha_1 N(t) e^{-\alpha_1 \int_0^t N(u)\,du} \sum_{k=0}^\infty p_k(t)\left(\int_0^t \rho(u)\pi(t-u)\,du\right)^k dt \\ &= \int_0^\infty \alpha_1 N(t) e^{-\alpha_1 \int_0^t N(u)\,du} \exp\left\{-\zeta\left(1 - \int_0^t \rho(u)\pi(t-u)\,du\right)\int_0^t N(u)\,du\right\} dt, \end{aligned}$$

where $\pi(t)$ and $\rho(t)$ are given by formulas (38) and (39), respectively.

The last formula assumes an especially simple form if the growth functions $N(t)$ and $M(t)$ are exponential with constant proliferation rates λ_1 and λ_2, respectively. In this case, introducing the notation

$$b(t) = \int_0^t \rho(u)\pi(t-u)\,du = \frac{\lambda_1 e^{\lambda_1 t}}{e^{\lambda_1 t} - 1}\int_0^t e^{-\lambda_1 u} e^{-\frac{\alpha_2}{\lambda_2}(e^{\lambda_2 u} - 1)}\,du,$$

we have

$$\Pi = \alpha_1 \int_0^\infty \exp\left\{\lambda_1 t - \frac{e^{\lambda_1 t} - 1}{\lambda_1}[\alpha_1 + \zeta(1 - b(t))]\right\} dt. \tag{40}$$

Incorporation of the distribution of time T of tumor onset into the above formulas presents no difficulties. Randomized versions of the model are also pretty straightforward.

We used formula (40) to conduct numerical experiments with the aim to study the behavior of $Pr\{W_1 < W_2\}$ as a function of model parameters.

Example 1. In this numerical experiment, the primary and the secondary (metastases) tumors grow at the same rate ($\lambda_1 = \lambda_2$) with the corresponding detection rate constants being equal ($\alpha_1 = \alpha_2$). Computations were carried out in accordance with formula (40) at the following values of model parameters: $\alpha_1 = \alpha_2 = 0.03, \lambda_1 = \lambda_2 = 3.0$. These values have no relevance to any actual

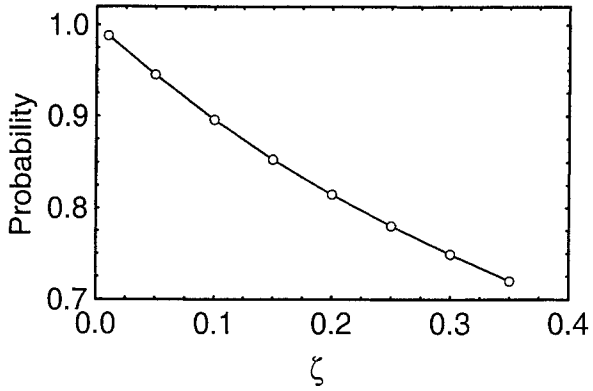


Fig. 8. The probability $\Pi = Pr\{W_1 < W_2\}$ as a function of ζ. The other parameter values are: $\alpha_1 = \alpha_2 = 0.03$, $\lambda_1 = \lambda_2 = 3.0$.

biological data and serve only as a purpose of illustration. The dependence of $\Pi = Pr\{W_1 < W_2\}$ on the parameter ζ is shown in Fig. 8. This dependence appears to be fairly flat in the given range of model parameters.

Example 2. Presented in Table 1 are the values of $\Pi = Pr\{W_1 < W_2\}$ as a function of the sensitivity parameter α_2 and the metastatic growth rate λ_2. The dependence of Π on the sensitivity parameter α_2 in the case of equal growth rates for the primary and metastases is shown in Fig. 9. All profiles of the dependencies under study indicate a low sensitivity of the probability Π to variations in the parameters of this model.

The probability Π can also be estimated non-parametrically by the corresponding relative frequency in cohort studies. Therefore, the usefulness of the above analytic formula for Π consists mainly in that the parametric estimate can be compared with its non-parametric counterpart, thereby suggesting an additional criterion for model validation. However, this criterion seems to be rather weak. Indeed, the results of the two numerical experiments suggest that the probability Π is relatively insensitive to the basic parameters incorporated into the model. In other words, inaccuracy of parameter estimates (obtained from data on age at diagnosis of the primary tumor and metastasis-free survival) does not appear to affect much the resultant estimate of the probability Π; this conclusion is very preliminary since it has been drawn from numerical experiments carried out in a limited range of parameter values.

Table 1
The probability $\Pi = Pr\{W_1 < W_2\}$ as a function of α_2 and λ_2 ($\alpha_1 = 0.03$, $\lambda_1 = 3.0$, $\zeta = 0.1$)

α_2	λ_2				
	2.0	2.5	3.05	3.5	4.0
0.015	0.966	0.956	0.941	0.924	0.903
0.03	0.936	0.919	0.896	0.874	0.848
0.06	0.885	0.861	0.831	0.804	0.775

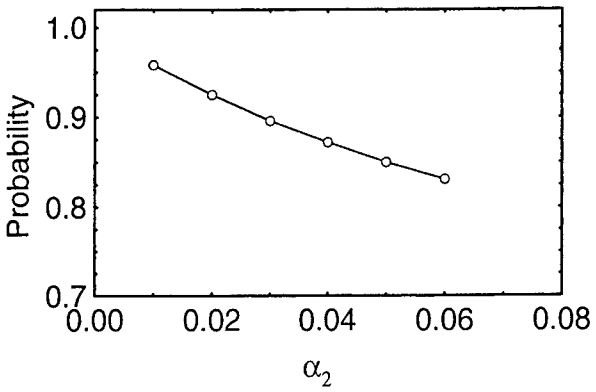


Fig. 9. The probability $\Pi = Pr\{W_1 < W_2\}$ as a function of α_2. The other parameter values are: $\alpha_1 = 0.03$, $\lambda_1 = \lambda_2 = 3.0$, $\zeta = 0.1$.

Acknowledgements

This research was supported by the NCI Grant 1U01 CA88177-01, Collaborative Linkage Grant under the NATO Science Programme, and Grant DAMD17-98-1-8256 awarded by the US Army Medical Research Acquisition Activity. A.T. and A.Yu.Y. are Huntsman Cancer Investigators. The analysis of epidemiological data was supported by the Utah Cancer Registry, which is funded by contract NO1-PC-67000 from the NCI with additional support from the Utah State Department of Health and the University of Utah. We would like to express our special gratitude to Professors V.V. Kalashnikov, E. von Collani and L.B. Klebanov for fruitful discussions. We are grateful to the reviewers whose comments have led to substantial improvements in the manuscript.

References

[1] T. Santner, D. Pearl, R. Bartoszyński, Dedication, Mathematical and Computer Modelling (special issue), in: A.Y. Yakovlev, S.H. Moolgavkar (Eds.), Stochastic Models and Data Analysis in Cancer Studies, in press.
[2] N.E. Atkinson, R. Bartoszyński, B.W. Brown, J.R. Thompson, On estimating the growth function of tumors, Math. Biosci. 67 (1983) 145.
[3] N.E. Atkinson, B.W. Brown, J.R. Thompson, On the lack of concordance between primary and secondary tumor growth rates, J. Nat. Cancer Inst. 78 (1987) 425.
[4] B.W. Brown, N.E. Atkinson, R. Bartoszyński, E.D. Montague, Estimation of human tumor growth rate from distribution of tumor size at detection, J. Nat. Cancer Inst. 72 (1984) 31.
[5] R. Bartoszyński, A modeling approach to metastatic progression of cancer, in: J.R. Thompson, B.W. Brown (Eds.), Cancer Modeling, Marcel Dekker, New York, 1987, p. 237.
[6] M. Klein, R. Bartoszyński, Estimation of growth and metastatic rates of primary breast cancer, in: O. Arino, D.E. Axelrod, M. Kimmel (Eds.), Mathematical Population Dynamics, Marcel Dekker, New York, 1991, p. 397.
[7] M. Kimmel, B.J. Flehinger, Nonparametric estimation of the size–metastasis relationship in solid cancers, Biometrics 47 (1991) 987.
[8] L.G. Hanin, A.D. Tsodikov, A.Y. Yakovlev, Optimal schedules of cancer surveillance and tumor size at detection, Math. Comput. Modelling, in press.

[9] P.S. Puri, J. Senturia, On a mathematical theory of quantal response assays, in: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4, University of California, Berkeley and Los Angeles, 1972, p. 231.
[10] P.S. Puri, A class of stochastic models of response after infection in the absence of defense mechanism, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4, University of California, Berkeley and Los Angeles, 1967, p. 511.
[11] P.S. Puri, A quantal response process associated with integrals of certain growth processes, in: M.T. Wasan (Ed.), Mathematical Aspects of Life Sciences, Queen's Papers in Pure and Applied Mathematics – No. 26, Queen's University, Kingston, Ont., Canada, 1971.
[12] H.M. Taylor, S. Karlin, An Introduction to Stochastic Modeling, Academic Press, San Diego, 1998.
[13] K.L. Chung, Lectures from Markov Processes to Brownian Motion, Springer, New York, 1982.
[14] T.E. Harris, The Theory of Branching Processes, Springer, Berlin, 1963.
[15] D.J. Daley, D. Vere-Jones, An Introduction to the Theory of Point Processes, Springer, New York, 1988.
[16] N.L. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, vol. 1, Wiley, New York, 1994.
[17] S.D. Grimshaw, Computing maximum likelihood estimates for the generalized Pareto distribution, Technometrics 35 (1993) 185.
[18] E.L. Lehmann, G. Casella, Theory of Point Estimation, Springer, New York, 1998.
[19] S. Zacks, The Theory of Statistical Inference, Wiley, New York, 1971.
[20] I.A. Ibragimov, R.Z. Has'minsky, Asymptotic Theory of Estimation, Nauka, Moscow, 1979 (in Russian).
[21] J.R.M. Hosking, J.R. Wallis, Parameter and quantile estimation for the generalized Pareto distribution, Technometrics 29 (1987) 339.
[22] S.T. Rachev, Probability Metrics and the Stability of Stochastic Models, Wiley, Chichester, UK, 1992.
[23] S.H. Moolgavkar, A.G. Knudson, Mutation and cancer: a model for human carcinogenesis, J. Nat. Cancer Inst. 66 (1981) 1037.
[24] S.H. Moolgavkar, E.G. Luebeck, Two-event model for carcinogenesis: biological, mathematical and statistical considerations, Risk Anal. 10 (1990) 323.
[25] S.H. Moolgavkar, D.J. Venzon, Two event model for carcinogenesis: incidence curves for childhood and adult tumors, Math. Biosci. 47 (1979) 55.
[26] W.F. Heidenreich, On the parameters of the clonal expansion model, Radiat. Environ. Biophys. 35 (1996) 127.
[27] L.G. Hanin, A.Yu. Yakovlev, A nonidentifiability aspect of the two-stage model of carcinogenesis, Risk Anal. 16 (1996) 711.
[28] W.F. Heidenreich, E.G. Luebeck, S.H. Moolgavkar, Some properties of the hazard function of the two-mutation clonal expansion model, Risk Anal. 17 (1997) 391.
[29] A. Kopp-Schneider, C.J. Portier, C.D. Sherman, The exact formula for tumor incidence in the two-stage model, Risk Anal. 14 (1994) 1079.
[30] Q. Zheng, On the exact hazard and survival functions of the MVK stochastic carcinogenesis model, Risk Anal. 14 (1994) 1081.
[31] E.G. Luebeck, W.F. Heidenreich, W.D. Hazelton, H.G. Paretzke, S.H. Moolgavkar, Biologically based analysis of the data for the Colorado uranium miners cohort: age, dose and dose-rate effects, Radiat. Res. 152 (1999) 339.
[32] C.J. Portier, A. Kopp-Schneider, C.D. Sherman, Calculating tumor incidence rates in stochastic models of carcinogenesis, Math. Biosci. 135 (1996) 129.
[33] D.W. Quinn, Calculating the hazard function and probability of tumor for cancer risk assessment when the parameters are time-dependent, Risk Anal. 9 (1989) 407.
[34] A.Yu. Yakovlev, E. Polig, A diversity of responses displayed by a stochastic model of radiation carcinogenesis allowing for cell death, Math. Biosci. 132 (1996) 1.
[35] L.G. Hanin, K.M. Boucher, Identifiability of parameters in the Yakovlev–Polig model of carcinogenesis, Math. Biosci. 160 (1999) 1.
[36] A. Yakovlev, A.D. Tsodikov, L. Bass, A stochastic model of hormesis, Math. Biosci. 116 (1993) 197.
[37] S.T. Rachev, C. Wu, A.Y. Yakovlev, A bivariate limiting distribution of tumor latency time, Math. Biosci. 127 (1995) 127.
[38] A.Yu. Yakovlev, A.D. Tsodikov, Stochastic Models of Tumor Latency and their Biostatistical Applications, World Scientific, Singapore, 1996.

[39] A.Y. Yakovlev, W.A. Müller, L.V. Pavlova, E. Polig, Do cells repair precancerous lesions induced by radiation?, Math. Biosci. 142 (1997) 107.
[40] I.L. Kruglikov, N.I. Pilipenko, A.D. Tsodikov, A.Y. Yakovlev, Assessing risk with doubly censored data: an application to the analysis of radiation-induced thyropathy, Statist. Prob. Lett. 32 (1997) 223.
[41] A.D. Tsodikov, W.A. Müller, Modeling carcinogenesis under a time-changing exposure, Math. Biosci. 152 (1998) 179.
[42] A.D. Tsodikov, F. Bruenger, R. Lloyd, E. Polig, S. Miller, A.Y. Yakovlev, Modeling and analysis of the latent period of osteosarcomas induced by incorporated ^{239}Pu: the role of immune responses, Math. Comput. Modell., in press.
[43] K. Boucher, A.Y. Yakovlev, Estimating the probability of initiated cell death prior to tumor induction, Proc. Nat. Acad. Sci. USA 94 (1997) 12776.
[44] A.D. Tsodikov, M. Loeffler, A.Y. Yakovlev, Assessing the risk of secondary leukemia in patients treated for Hodgkin's disease. A report from the International Database on Hodgkin's disease, J. Biol. Syst. 5 (1997) 433.
[45] A.Y. Yakovlev, L.V. Pavlova, Mechanistic modeling of multiple tumorigenesis: an application to data on lung tumors in mice exposed to urethane, Ann. NY Acad. Sci. 837 (1997) 462.
[46] A.D. Tsodikov, M. Loeffler, A.Y. Yakovlev, A parametric regression model with time dependent covariates: an application to the analysis of secondary leukemia. A report from the International Database on Hodgkin's disease, Stat. Med. 17 (1998) 27.
[47] K. Boucher, L.V. Pavlova, A.Yu. Yakovlev, A model of multiple tumorigenesis allowing for cell death: quantitative insight into biological effects of urethane, Math. Biosci. 150 (1998) 63.
[48] K.M. Boucher, R.A. Kerber, The shape of the hazard function for cancer incidence, Math. Comput. Modell., in press.
[49] V.E. Johnson, J.H. Albert, Ordinal Data Modeling, Springer, New York, 1999.
[50] D.M. Himmelblau, Applied Nonlinear Programming, McGraw-Hill, New York, 1972.
[51] J.D. Kalbfleisch, R.L. Prentice, The Statistical Analysis of Failure Time Data, Wiley, New York, 1980.
[52] A.Yu. Yakovlev, L.G. Hanin, S.T. Rachev, A.D. Tsodikov, Distribution of tumor size at detection and its limiting form, Proc. Nat. Acad. Sci. USA 93 (1996) 6671.
[53] L.G. Hanin, S.T. Rachev, A.D. Tsodikov, A.Yu. Yakovlev, A stochastic model of carcinogenesis and tumor size at detection, Adv. Appl. Prob. 29 (1997) 607.
[54] D.R. Cox, P.A.W. Lewis, The Statistical Analysis of Series of Events, Methuen, London; Wiley, New York, 1966.
[55] A.D. Tsodikov, B. Asselain, A.Y. Yakovlev, A distribution of tumor size at detection: an application to breast cancer data, Biometrics 53 (1997) 1495.
[56] E.D. Yorke, Z. Fuks, L. Norton, W. Whitmore, C.C. Ling, Modeling the evelopment of metastases from primary and local recurrent tumors: comparison with a clinical data base for prostatic cancer, Cancer Res. 53 (1993) 2987.
[57] D.R. Cox, V. Isham, Point Processes, Chapman and Hall, London, 1980.

PERGAMON Mathematical and Computer Modelling 33 (2001) 1361–1375

MATHEMATICAL AND COMPUTER MODELLING

www.elsevier.nl/locate/mcm

The Shape of the Hazard Function for Cancer Incidence

K. M. BOUCHER* AND R. A. KERBER
Huntsman Cancer Institute and Department of Oncological Sciences
University of Utah, 2000 East North Campus Drive
Salt Lake City, UT 84112, U.S.A.
ken.boucher@hci.utah.edu

Abstract—A population-based cohort consisting of 126,141 men and 122,208 women born between 1874 and 1931 and at risk for breast or colorectal cancer after 1965 was identified by linking the Utah Population Data Base and the Utah Cancer Registry. The hazard function for cancer incidence is estimated from left truncated and right censored data based on the conditional likelihood. Four estimation procedures based on the conditional likelihood are used to estimate the age-specific hazard function from the data; these were the life-table method, a kernel method based on the Nelson Aalen estimator, a spline estimate, and a proportional hazards estimate based on splines with birth year as sole covariate.

The results are consistent with an increasing hazard for both breast and colorectal cancer through age 85 or 90. After age 85 or 90, the hazard function for female breast and colorectal cancer may reach a plateau or decrease, although the hazard function for male colorectal cancer appears to continue to rise through age 105. The hazard function for both breast and colorectal cancer appears to be higher for more recent birth cohorts, with a more pronounced birth-cohort effect for breast cancer than for colorectal cancer. The age specific hazard for colorectal cancer appears to be higher for men than for women. The shape of the hazard function for both breast and colorectal cancer appear to be consistent with a two-stage model for spontaneous carcinogenesis in which the initiation rate is constant or increasing. Inheritance of initiated cells appears to play a minor role.

Keywords—Hazard function, Truncation, Survival analysis, Breast cancer, Colorectal cancer.

1. INTRODUCTION

The shape of the hazard function may lead to insights into the biology of carcinogenesis which may not be easily discernable from a study of the survival function alone. For example, it is typical in the analysis of tumor recurrence data to find a hazard function that is bimodal or unimodal, and that tends to zero as time tends to infinity [1]. The modes of the hazard may be interpreted biologically as arising from two different types of failure, one that tends to occur earlier and one that tends to occur later. The decrease in the hazard function to zero may lead

This research was supported, in part, by NCI Cancer Center Support Grants 5P30 CA 4201 and 2P30 CA 42014, U.S. Army Medical Research and Material Command Grant DAMD17-98-1-8256, and by NIH/NCI Grant R29 CA69421. Partial support was provided by the Huntsman Cancer Institute for the Utah Population Data Base. The Utah Cancer Registry was supported by NCI Grant NO1 CN 6700.
In addition, the authors would like to thank A. Y. Yakovlev for many helpful discussions.
*Author to whom all correspondence should be addressed.

0895-7177/01/$ - see front matter Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX
PII: S0895-7177(00)00321-6

one to conclude that there is a nonzero cured fraction. In fact, if we let $\lambda(t)$ denote the hazard function, and p the probability of cure, it follows from the formula

$$p = \lim_{t\to\infty} \exp\left\{-\int_0^t \lambda(u)\,du\right\}$$

that there are individuals who have been "cured" in the population exactly when the hazard function has finite integral. In particular, $\lim_{t\to\infty} \lambda(t) = 0$, provided the limit exists.

If the hazard function under study is from disease incidence, the "cured fraction" must be reinterpreted as the fraction of the population that is "immune" to the disease. If the cumulative hazard appears to be bounded, for example, one should expect the existence of a nonzero immune fraction. More generally, a large degree of heterogeneity in disease susceptibility may lead to a population hazard function with one or more well-defined maxima. The maxima may correspond to discrete subpopulations with different genetic predisposition to disease. A maximum may also result from a continuous frailty, as the surviving population at higher ages may be overrepresented by individuals with lower risk [2].

Both breast and colorectal cancer are syndromes in which an inherited susceptibility has been shown to play a role. Inherited mutations in p53, BRCA1, BRCA2, the ataxia-telangiectasia gene (AT), HRAS, and the androgen receptor gene (AR) have been shown to play a role in breast cancer susceptibility [3]. About 56% of carriers of the mutation BRCA1 or BRCA2 will get breast cancer by the age of 70 years [4]. BRCA1 has an estimated allele frequency of between 0.0002 and 0.001 (95% CI) [5], and accounts for about 3% of diagnosed breast cancer [6]. The allele frequency of mutations in BRCA2 is estimated at 0.00022 [7]. Germline mutations in p53 and AR are extremely rare, and mutations in the HRAS1 minisatellite locus which confer increased risk of breast cancer are also rare, having an estimated population frequency of 6% [3]. In a study of 100 Finnish breast cancer families analyzed by protein truncation tests and direct sequencing, Vehmanen *et al.* [8] found that only 21% of breast cancer families were accounted for by mutations of BRCA1 and BRCA2, providing indirect evidence for the existence of other, undiscovered breast cancer genes.

Indirect evidence also exists for the existence of additional colorectal cancer genes. Inherited mutations in polyposis coli (APC) gene and the hereditary nonpolyposis colon cancer syndrome (HNPCC) genes hMSH2, and hMLH1 have been shown to play a role in colon cancer susceptibility [3]. After segregation analysis of 203 pedigrees, Houlston *et al.* [9] concluded that dominant colorectal cancer genes with a frequency of 0.006 account for an estimated 81% of colorectal cancers in patients under 35, 59% in patients between 35 and 49, decreasing to 16% in patients over 65. The I1307K mutation of the APC gene, found in Ashkenazi Jews, confers an estimated relative risk of 1.7 for colorectal cancer (95% CI 1.01–2.87) [10]. APC and HNPCC are rare, and contribute to a small percentage of colorectal cancer cases [3].

Additional insight can be gleaned from the hazard function for cancer incidence in the framework of a mechanistic model of carcinogenesis. The most widely accepted model is the Moolgavkar-Venzon-Knudson two-stage clonal expansion model [11,12]. The Moolgavkar-Venzon-Knudson model has the following assumptions.

ASSUMPTION A. Normal, susceptible target cells are initiated according to a (nonhomogeneous) Poisson process with intensity $\nu(t)$.

ASSUMPTION B. The expansion of the colony of initiated cells and malignant transformation is specified by a stochastic birth-death-migration process with the division, death (or differentiation) and transformation. Premalignant cells either divide into two premalignant cells with rate $\alpha(t)$, die with rate $\beta(t)$, or divide asymmetrically into one premalignant cell and one malignant cell with rate $\mu(t)$.

It has been shown that the hazard function for the Moolgavkar-Venzon-Knudson model with constant parameters increases monotonically and approaches an asymptote [13]. An asymptotic

value for the hazard is also reached for the Moolgavkar-Venzon-Knudson model with piecewise constant parameters, and in that case, the value of the asymptote depends only on the value of the coefficients in the unbounded interval [13,14].

Expressions for the survivor function were first obtained by Moolgavkar and Luebeck [13]. A simple explicit formula for the survivor function $S(t)$ for the Moolgavkar-Venzon-Knudson model with constant parameters was obtained by Kopp-Schneider *et al.* [15] and Zheng [16],

$$S(t) = \left[\frac{2ce^{0.5(-\alpha+\beta+\mu-c)t}}{(-\alpha+\beta+\mu+c)+(\alpha-\beta-\mu+c)e^{-ct}}\right]^{\nu/\alpha}, \tag{1}$$

where $c = \sqrt{(\alpha+\beta+\mu)^2 - 4\alpha\beta}$. Zheng also presented an expression for the probability generating function for the number of malignant cells given a single malignant cell at time $t = 0$, allowing an expression for the promotion time distribution

$$F(t) = \frac{(\alpha-\beta-\mu+c)(\alpha-\beta-\mu-c)e^{-ct} + (\alpha-\beta-\mu+c)(-\alpha+\beta+\mu+c)}{2\alpha\left[(\alpha-\beta-\mu+c)e^{-ct} + (-\alpha+\beta+\mu+c)\right]} \tag{2}$$

to be given. It is easy to see that $S(t)$ and $F(t)$ above are related by the formula

$$S(t) = \exp\left\{-\nu\int_0^t F(x)\,dx\right\}, \tag{3}$$

which was shown by Hanin and Yakovlev [17] to be valid in a more general setting.

Yakovlev and Tsodikov [18] replace Assumption B above with the following assumption.

ASSUMPTION C. Progenitor cells are transformed into malignant lesions at a random with cumulative distribution function $F(x)$. All progenitor cells are promoted independently of one another.

Assuming $F(0) = 0$, it follows that the process of malignant transformation is also a Poisson process, with integral rate $\Lambda(t) = \int_0^t \nu(u)F(t-u)\,du$. As in the Moolgavkar-Venzon-Knudson model, the simplest model of spontaneous carcinogenesis takes $\nu(t) = \nu$ to be constant, in which case $\Lambda(t) = \nu\int_0^t F(u)\,du$ and the hazard function for time-to-tumor, given by $\lambda(t) = \nu F(t)$, is nondecreasing. The probability $S(t)$ that there are no malignancies by time t is then given by (3).

This model may easily be modified to handle inherited lesions, via the limiting case where ν is taken to be a delta function at the origin. If $F(t)$ is assumed to be absolutely continuous, then the integral rate $\Lambda(t)$ is equal to $\nu F(t)$ and the hazard function $\lambda(t) = F'(t) = f(t)$, where $f(t)$ is the density function associated with $F(t)$. We see that the hazard function for spontaneous and inherited lesions are quite likely to have very different shapes.

Even though a thorough study of the hazard function may lead to new insight into the process of carcinogenesis, few if any population-based cohorts have been analyzed to determine the hazard function for cancer incidence. In addition, time-dependent variation in environmental risk factors for cancer may cause estimates from a cross-sectional study to be misleading. In this paper, the age specific hazard function for both breast and colorectal cancer incidence are estimated using data from the Utah Cancer Registry and the Utah Population Data Base. We see that the hazard function for both these types of cancer appears to be increasing monotonically, at least through age 85 or 90. In the context of the above mechanistic models of carcinogenesis, we will see that risks for both these cancers at the population level appear to be relatively homogeneous, with negligible inherited component.

2. METHODS

2.1. Data

The data for this study was obtained by linking records from the Utah Population Data Base (UPDB) with the Utah Cancer Registry (UCR). The UPDB consists of the genealogical records of more than 1,000,000 individuals who were born, died, or married in Utah, or en route to Utah

during the nineteenth and twentieth centuries. Since 1973 the UCR has been reporting to the National Cancer Institutes Surveillance Epidemiology and End Results (SEER) program, and is required to maintain very high standards for case reporting and follow-up, and to periodically undergo quality control audits by SEER personnel to assure uniformly high quality and consistency from year to year. The available follow-up information comes either from Utah death certificates, which have been linked to the UPDB genealogical data every year from 1933 through the beginning of 1997, or from linkage of the HCFA beneficiary data to the UPDB. The study population consisted of 126,141 men and 122,208 women recorded in the Utah Population Database, who were born from 1874 to 1931 and for whom follow-up information is available that places them in Utah during the years of operation of the Utah Cancer Registry (1966–present). Subjects with purported follow-up past age 105 were excluded from the data. There are 5,372 cases of female breast cancer and 5,177 cases of colorectal cancer represented in the data. Analyses were performed on subcohorts based on birth year (1874–1889, 1890–1899, 1900–1909, 1910–1919, and 1920–1931) and gender. For each gender, the entire cohort (birth years 1874–1931) was also analyzed as a whole. The total number of subjects and cases of breast and colorectal cancer for each birth subcohort and gender are given in Tables 1 and 2. Male breast cancer was not analyzed.

Table 1. Number of female subjects and cases of breast and colorectal cancer, stratified by birth year.

Birth Years	Number of Subjects	No. of Breast Cancer Cases	No. of Colorectal Cancer Cases
1874–1889	10,115	145	116
1890–1899	19,352	564	435
1900–1909	27,138	1258	755
1910–1919	31,162	1709	752
1920–1931	34,441	1696	448
Total	122,208	5372	2106

Table 2. Number of male subjects and cases of colorectal cancer, stratified by birth year.

Birth Years	Number of Subjects	No. of Colorectal Cancer Cases
1874–1889	6,850	101
1890–1899	16,307	341
1900–1909	27,122	768
1910–1919	34,731	874
1920–1931	41,131	587
Total	126,141	2671

2.2. Truncation: Nonparametric Estimation

We wish to estimate the age specific hazard function for breast and colorectal cancer from the data described above, taking into account that the data is subject to random truncation: cases which occurred during or before 1965 are not recorded in the dataset. Subjects were between the ages of 34 and 86, at the time of truncation. Thus, analysis of the data must take into account not

only to the effects of right censoring, but also the effects of left truncation due to delayed entry into the risk set. The topic of random truncation is not mentioned in several authoritative texts such as [19] and [20], and may be unfamiliar to some readers, and therefore, will be discussed in this and the following section.

Let the truncation time Y have distribution function $G(y)$ and the failure time (time of cancer diagnosis) X have distribution function $F(x)$. We require that truncation be independent of failure and for simplicity assume no censoring for the present. Observations are conditional on $X > Y$. Let $G^*(y)$ and $F^*(x)$ be the corresponding distribution functions, conditional on $X > Y$. Let $S(x) = 1 - F(x)$ be the survivor function of X. Suppose that we have observations $(Y_1^*, X_1^*), \ldots, (Y_n^*, X_n^*)$ from the conditional distribution. The full likelihood of the observed data is given by

$$L = \prod_{j=1}^{n} \left[\frac{dF(X_j)\, dG(Y_j)}{\alpha} \right], \tag{4}$$

where $\alpha = \iint_{y \le x} dF(x)\, dG(y)$. A key observation is that if X and Y are independent, then the hazard of X given $X > Y = y$ at $x > y$ is equal to the hazard of X at x [21,22]. This observation leads to the result, first mentioned by Kaplan and Meier [23], that if the distribution $G(t)$ is allowed to vary freely, the natural generalization of the product limit estimator, given by the formula

$$\hat{S}(t) = \prod_{X_i^* \le t} \left(1 - \frac{1}{R(X_i^*)} \right), \tag{5}$$

where $R(u) = \#\{Y_i^* < U \le X_i^*\}$ is the number at risk at U, is the nonparametric maximum likelihood estimator (NPMLE) of the survivor function $S(t)$ of X (see, for example, [21,22,24]).

This result extends naturally to the case with random independent censoring [24]. It also easily follows that in the nonparametric setting (again with no censoring), maximizing (4) is equivalent to maximizing the conditional likelihood of $(X_1^*, \ldots, X_n^*)$ given $(Y_1^*, \ldots, Y_n^*)$, which can be written

$$CL = \prod_{i=1}^{n} \frac{f(X_i^*)}{S(Y_i^*)} \tag{6}$$

(see, for example, [23–26]). Maximizing the conditional likelihood also leads to the familiar Nelson-Aalen estimator for the integrated hazard function $H(t)$ of X [24], which is given by

$$\hat{\Lambda}(t) = \sum_{X_i^* \le t} R(X_i^*)^{-1}. \tag{7}$$

These results can be extended to the case of right censoring [24].

2.3. Truncation: Parametric Models

We consider the situation where X and Y are independent, $F(x)$ is parametrized, while $G(y)$ is allowed to vary freely. In a later section, $F(x)$ will be come from a quadratic spline model.

The data are independent pairs $(y_1, x_1), \ldots, (y_n, x_n)$ from the joint distribution (Y, X), conditional on $(Y < X)$. We suppose, for simplicity, that there are no ties among $y_1, y_2, \ldots, y_n$, and suppose X has absolutely continuous distribution function coming from a family $F(x; \vec{z})$ parameterized by a vector $\vec{z}$, with corresponding survival function $S(x; \vec{z}) = 1 - F(x; \vec{z})$ and density $f(x; \vec{z})$. The NPMLE for G should consist of (unknown) point masses $q_1, q_2, \ldots, q_n$ placed at the points $y_1, y_2, \ldots, y_n$. The logarithm of the complete likelihood (4) can be rewritten

$$\log(L) = \sum_{i=1}^{n} \left[\log\left(f(x_i; \vec{z})\right) + \log(q_i) \right] - n \log \left[\sum_{j=1}^{n} S(y_i; \vec{z})\, q_i \right]. \tag{8}$$

If we factor the out the part of the likelihood corresponding to (6), the logarithm is given by

$$\log(CL) = \sum_{i=1}^{n} \left[\log\left(f\left(x_i; \vec{z}\right)\right) - \log\left(S\left(y_i; \vec{z}\right)\right)\right]. \tag{9}$$

We now discuss the changes which must be made when censoring and additional covariates are present. If $\vec{s}$ is a vector of additional covariates, $\lambda(x, \vec{s}; \vec{z})$ denotes the hazard associated with $F(x, \vec{s}; \vec{z})$ and $\Lambda(x, \vec{s}; \vec{z})$ the cumulative hazard, we note that (9) becomes

$$\log(CL) = \sum_{i=1}^{n} \left[\log\left(\lambda\left(x_i, \vec{s}_i; \vec{z}\right)\right) - \left(\Lambda\left(x_i, \vec{s}_i; \vec{z}\right) - \Lambda\left(y_i, \vec{s}_i : \vec{z}\right)\right)\right]. \tag{10}$$

In the presence of right censoring which is independent of both the failure and truncation times, x_i is replaced in the above formulation by the minimum of the failure and censoring time. The term $f(x, \vec{s}; \vec{z})$ in the likelihood is replaced by $f(x, \vec{s}; \vec{z})^{\delta} S(x, \vec{s}; \vec{z})^{1-\delta}$, where $\delta_i = 1$ if observation i is a failure and $\delta_i = 0$ otherwise, and the conditional likelihood (6) (with x_i, $\vec{s}_i$, and y_i regarded as fixed) becomes

$$CL = \prod_{i=1}^{n} \frac{\left[f\left(x_i, \vec{s}_i; \vec{z}\right)^{\delta} S\left(x_i, \vec{s}_i; \vec{z}\right)^{(1-\delta)}\right]}{S\left(y_i, \vec{s}_i; \vec{z}\right)}.$$

In this setting, $\log(CL)$ becomes

$$\log(CL) = \sum_{i=1}^{n} \left[\delta_i \log\left(\lambda\left(x_i, \vec{s}_i; \vec{z}\right)\right) - \left(\Lambda\left(x_i, \vec{s}_i; \vec{z}\right) - \Lambda\left(y_i, \vec{s}_i; \vec{z}\right)\right)\right]. \tag{11}$$

In the subsequent analysis, we choose to maximize (11) rather than the full likelihood. In a separate paper, we will show that this is equivalent to maximizing the full likelihood, under appropriate conditions.

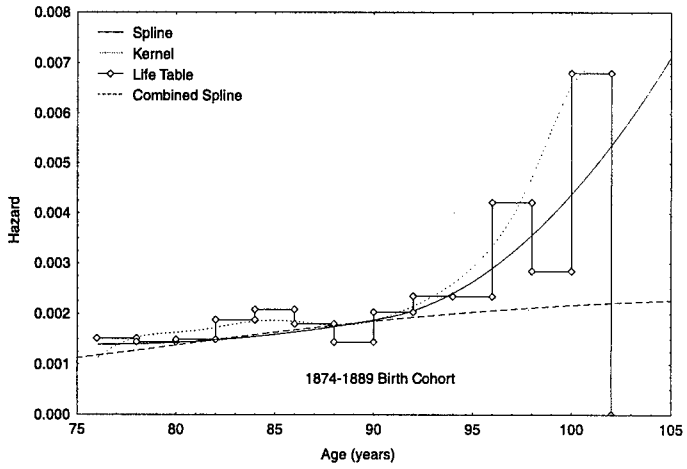
2.4. Spline Models

We choose to model the hazard via quadratic splines as in [27]. A quadratic spline with m knots specifies the hazard to be of the form

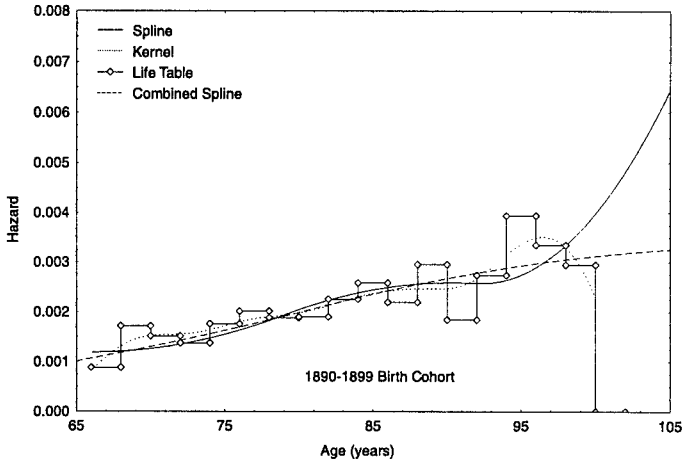
$$\lambda_m(t) = \sum_{i=0}^{2} \gamma_{0i} t^k + \sum_{j=1}^{m} \gamma_{j2} \left(t - \tau_j\right)_+^2, \tag{12}$$

where $(x)_+ = \max(x, 0)$. For each birth cohort, we fit splines with knots which were equally spaced in the interior of the interior $[T_{\min}, T_{\max}]$, where $T_{\min}$ is the minimum truncation age in the cohort and $T_{\max}$ the maximum follow-up (failure or censoring) time. Restrictions were placed on the coefficients to ensure that $\lambda_m(t)$ remained positive for all t. Thus, with m knots the number of parameters was $m + 3$. Models were fit using maximum likelihood techniques applied to the conditional likelihood, as given by (11).

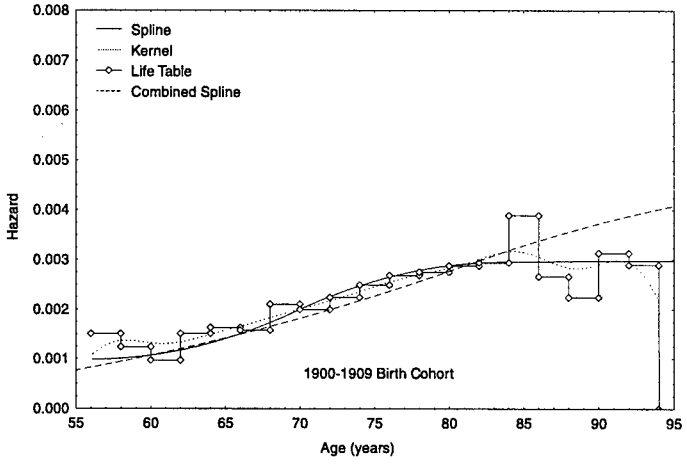
The hazard function was estimated for breast cancer incidence (women only) and for colorectal cancer incidence (both men and women). The spline estimates were computed by maximizing $\log(CL)$ using the algorithm of Powell [28]. We started with one knot and increased the number of knots until the fit was not improved, as determined by the likelihood ratio test at the significance level $\alpha = 0.05$. Two other subcohort estimates of the hazard function were computed for comparison with the spline estimator; a life table version of (5), and a Gaussian kernel estimate based on the Nelson-Aalen estimator (7).



(a)

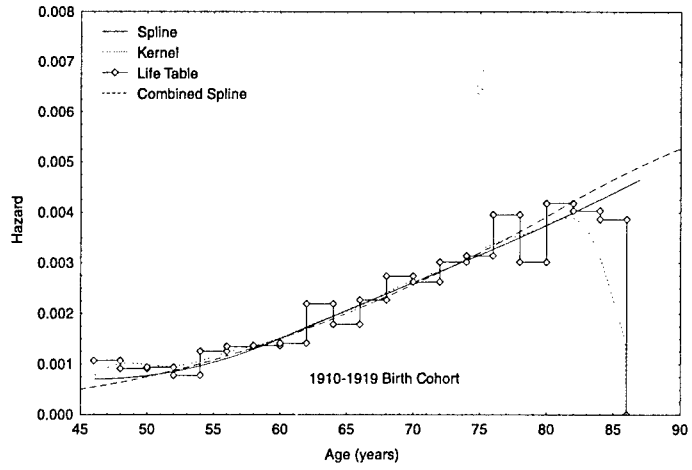


(b)

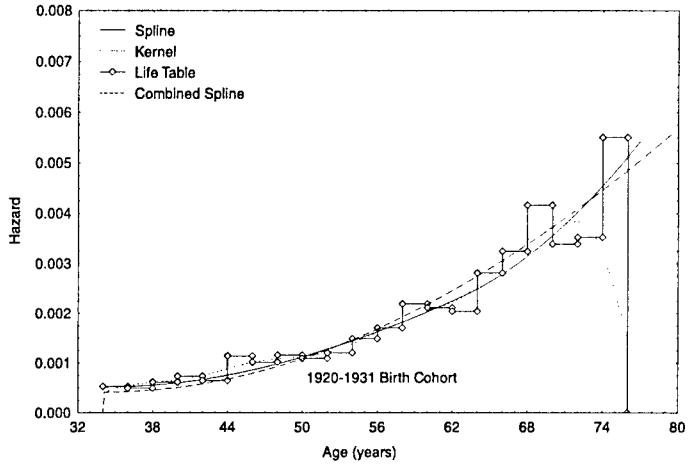


(c)

Figure 1. Four estimates of the age-specific hazard function for female breast cancer, stratified by birth cohort: a spline estimate (labeled "Spline"), a kernel estimate based on the Nelson-Aalen estimator (labeled "Kernel"), a life table estimate (labeled "Life Table"), and a proportional hazards spline estimate using all strata, with birth year as sole covariate, set at the stratum mean (labeled "Combined Spline").



(d)



(e)

Figure 1. (cont.)

2.5. Proportional Hazards

It became clear, when fitting models to the subcohorts, that there was a birth cohort effect in the data. At the same time, we wished to have estimates of the hazard for the entire age range of 34–100+ years. We therefore fit proportional hazards models with splines $\lambda_m(t)$ for the baseline hazard and a single covariate s representing birth year. The resulting hazard function has the form

$$\lambda_m(t, s; \beta) = \exp(\beta s)\lambda_m(t). \tag{13}$$

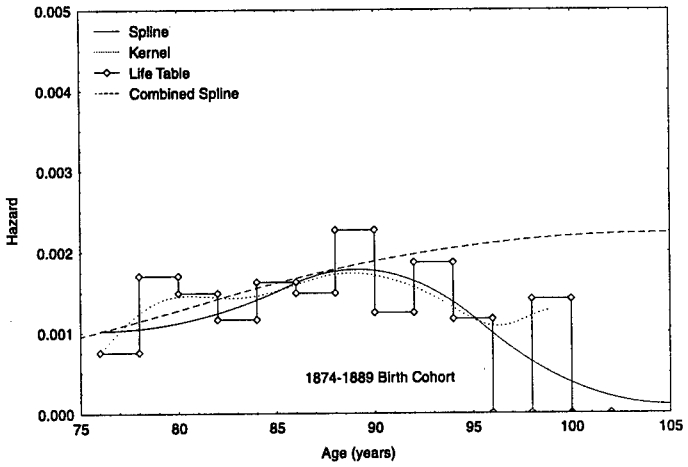
The model was again fit using the conditional likelihood of the form

$$\log(CL) = \sum_{i=1}^{n} \left[\delta_i \log\left(\lambda_m\left(x_i, s_i; \beta\right)\right) - \left(\Lambda\left(x_i, s_i; \beta\right) - \Lambda\left(y_i, s_i; \beta\right)\right)\right], \tag{14}$$

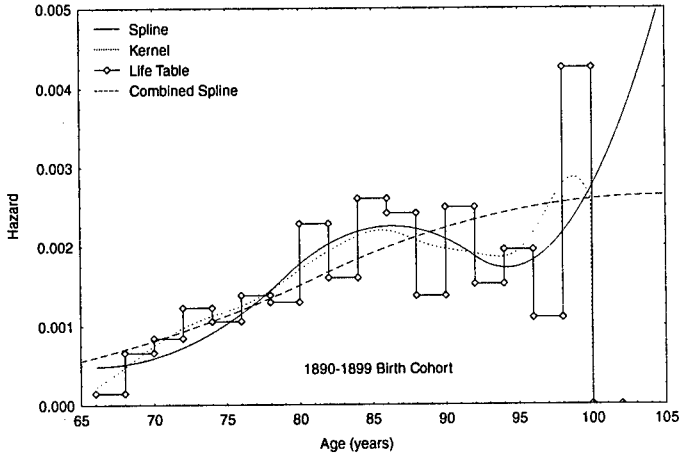
which is (11) with $\lambda(x, \vec{s}, \vec{z}) = \lambda_m(x_i, s_i; \beta)$.

3. RESULTS

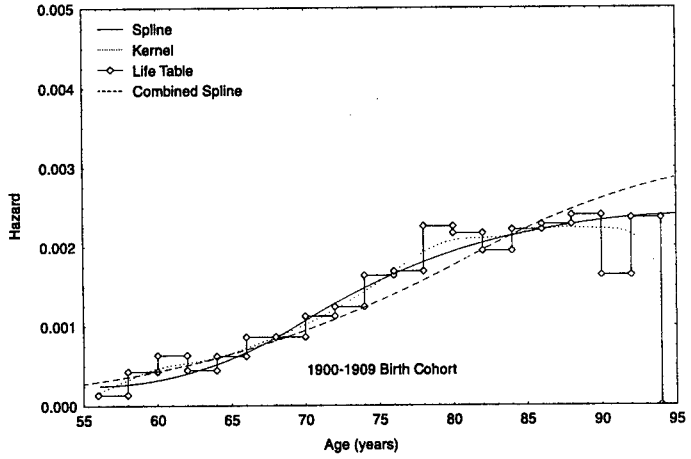
Estimates of the age specific hazard for female breast cancer are presented in Figure 1 for the 1874–1889, 1890–1899, 1900–1909, 1910–1919, and 1920–1931 birth subcohorts. Age specific



(a)

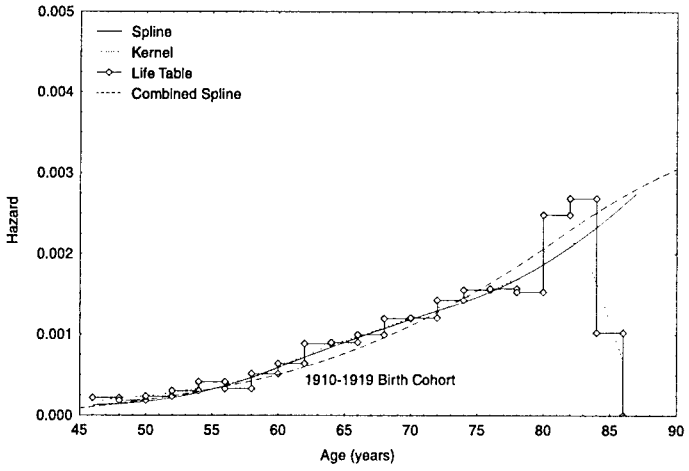


(b)

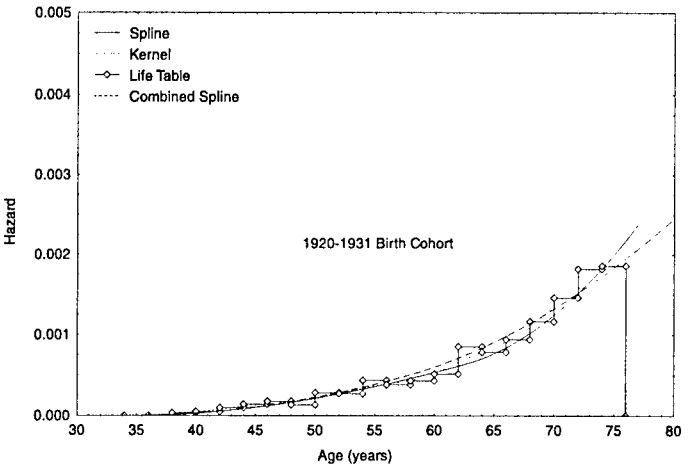


(c)

Figure 2. Four estimates of the age-specific hazard function for female colorectal cancer, stratified by birth cohort: a spline estimate (labeled "Spline"), a kernel estimate based on the Nelson-Aalen estimator (labeled "Kernel"), a life table estimate (labeled "Life Table"), and a proportional hazards spline estimate using all strata, with birth year as sole covariate, set at the stratum mean (labeled "Combined Spline").



(d)

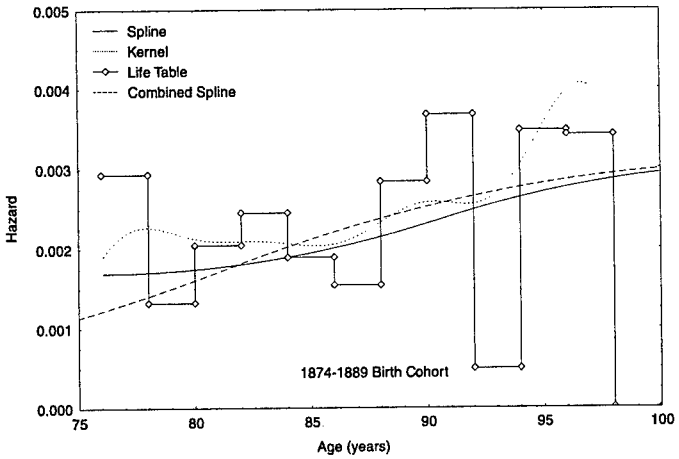


(e)

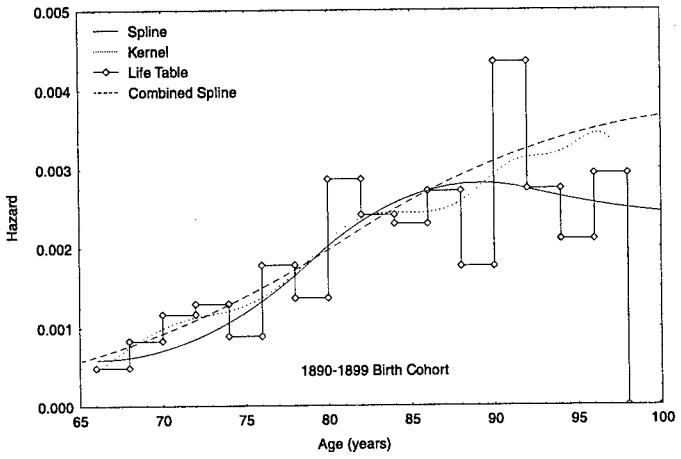
Figure 2. (cont.)

hazards for colorectal cancer are presented in Figures 2 and 3, stratified by birth cohort and gender. Each figure presents three estimates of the hazard from the subcohort alone, namely the life table estimate, the kernel estimate based on the Nelson-Aalen estimator and a spline estimate, as well as one gender-specific estimate from a proportional hazards model with birth year as covariate, fit to data from all birth subcohorts (1874–1931). The covariate is set to the mean birth year of the subcohort. We note that approximately 40 years of follow-up are available for any one subcohort, as follow-up data are available from approximately 1965–1995.

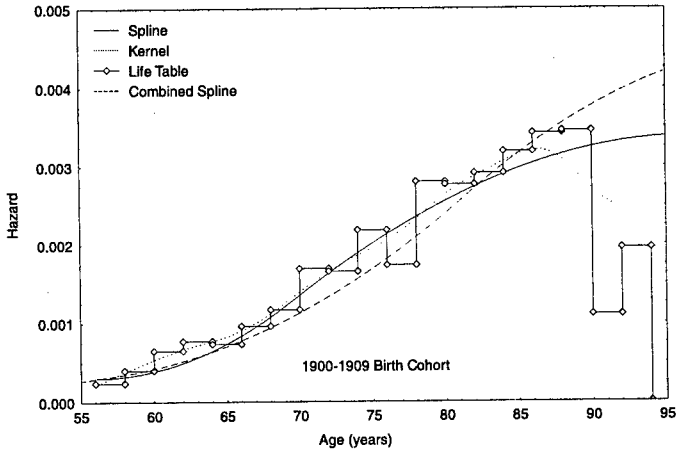
We found that splines with very few knots appeared to fit the data. In all but one case, two knots were sufficient for the spline estimates, as determined by the likelihood ratio test, and in the remaining case (breast cancer, birth years 1874–1889), one knot sufficed. The hazard function for both breast and colorectal cancer appears to increase monotonically, at least until the age of 85 or 90, when the subcohort specific estimates of the hazard estimates for women for both breast and colon cancer appear to flatten or decrease while the estimate for men appears to continue to increase. (In each of the three cases, the proportional hazards model provides estimates of the hazard function which increase through all ages.) We also note that, in all the proportional hazards models, the birth cohort effect was highly significant ($p < 0.0001$). We also see from the subcohort analysis that the proportional hazards assumption appears to be adequate, at least up until the age of 85 or 90, when proportionality may fail for women.



(a)

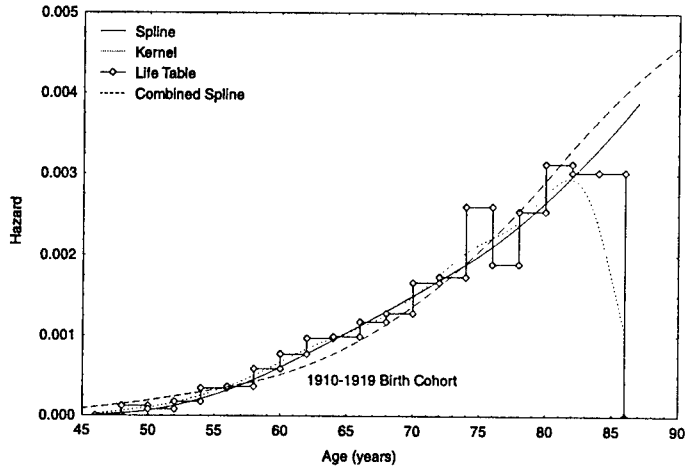


(b)

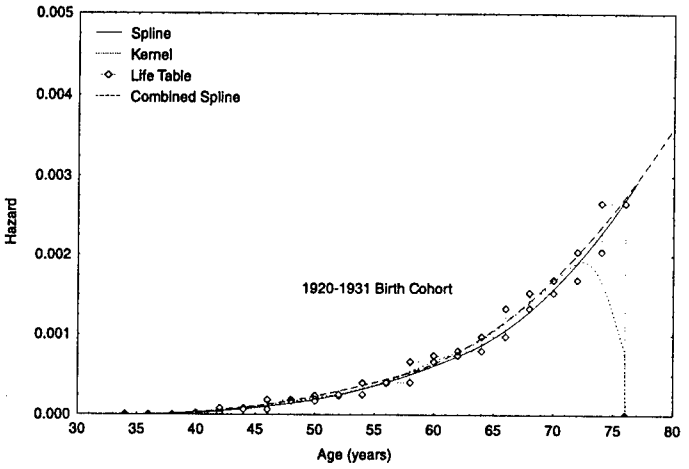


(c)

Figure 3. Four estimates of the age-specific hazard function for male colorectal cancer, stratified by birth cohort: a spline estimate (labeled "Spline"), a kernel estimate based on the Nelson-Aalen estimator (labeled "Kernel"), and a life table estimate (labeled "Life Table"), and a proportional hazards spline estimate using all strata, with birth year as sole covariate, set at the stratum mean (labeled "Combined Spline").



(d)



(e)

Figure 3. (cont.)

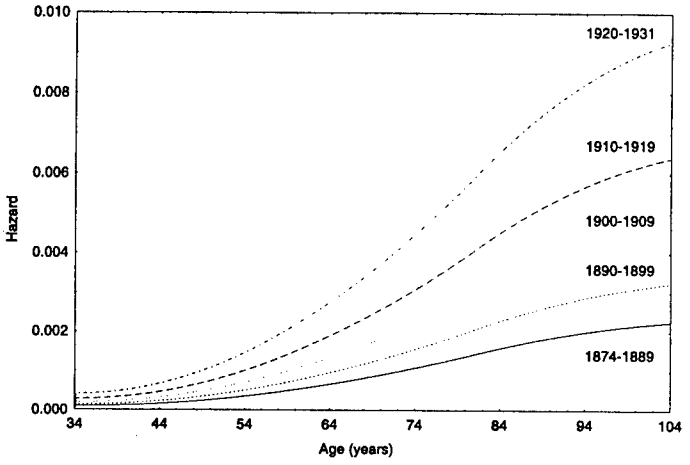


Figure 4. Comparison of the age-specific hazard function estimates for female breast cancer for various birth cohort strata from a proportional hazards model spline model. Birth year covariate set at the mean value for each stratum: 1884.41 for the 1874–1889 stratum, 1894.90 for the 1890–1899 stratum, 1904.54 for the 1900–1909 stratum, 1914.52 for the 1910–1919 stratum, and 1925.24 for the 1920–1931 stratum.

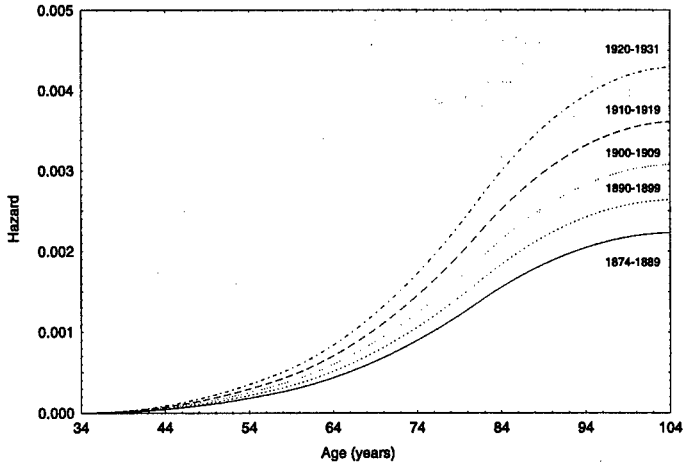


Figure 5. Comparison of the age-specific hazard function estimates for female colorectal cancer for various birth cohort strata from a proportional hazards model spline model. Birth year covariate set at the mean value in each stratum: 1884.41 for the 1874–1889 stratum, 1894.90 for the 1890–1899 stratum, 1904.54 for the 1900–1909 stratum, 1914.52 for the 1910–1919 stratum, and 1925.24 for the 1920–1931 stratum.

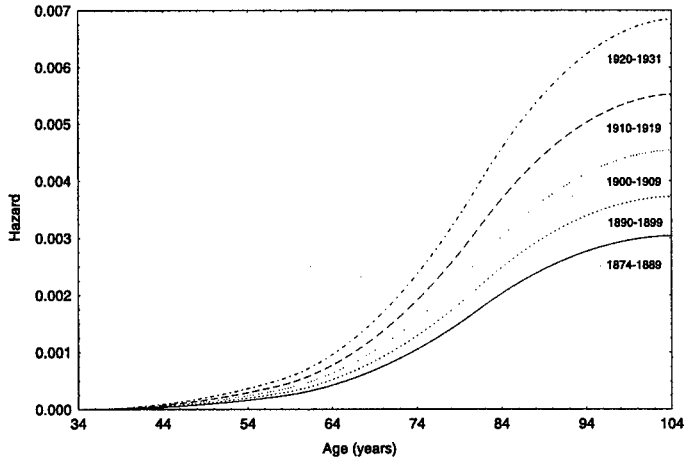


Figure 6. Comparison of the age-specific hazard function estimates for male colorectal cancer for various birth cohort strata from a proportional hazards model spline model. Birth year covariate set at the mean value in each stratum: 1884.74 for the 1874–1889 stratum, 1895.06 for the 1890–1899 stratum, 1904.74 for the 1900–1909 stratum, 1914.57 for the 1910–1919 stratum, and 1925.31 for the 1920–1931 stratum.

We also note that the colorectal cancer risk estimates are higher for men than for women. For example, the estimated age specific yearly hazard for the 1920–1931 birth cohort at age 70 is approximately 0.0013 for women, and about 0.0017 for men, or about 30% higher for men.

The estimated hazards from the proportional hazards models over a seventy year range are presented in Figures 4–6. The estimated hazards increase as the birth cohorts become more recent, with coefficient estimates of $\beta = 0.0347$ (year^{-1}) for female breast cancer, $\beta = 0.016$ (year^{-1}) for female colorectal cancer, and $\beta = 0.020$ (year^{-1}) for male colorectal cancer. Thus, the additional hazard for more recent birth cohorts appears to be more pronounced for breast cancer than for colorectal cancer.

4. DISCUSSION

As noted in the introduction, the presence of a large degree of heterogeneity in the risk for a population may lead to a decreasing age specific hazard function. Since we see little or no

evidence of a decreasing hazard for either breast or colorectal cancer at least until age 85 or 90, it appears that the risk is relatively homogeneous for both these cancers over this age range. In particular, there appears to be little evidence for a high immune fraction for either breast or colorectal cancer. We should also note that the presence of a monotone increasing hazard over a limited range does not completely rule out heterogeneity. The data is quite consistent with the degree of heterogeneity that might result from known cancer genes, as long as the risk is generally increasing (at least through age 90) in the population as a whole. There is little or no evidence of an inherited component to the risk, as a large inherited component might be expected to provide a local maxima to the hazard rather early in life, certainly prior to age 85.

One may extend the more general two-stage model of carcinogenesis presented in the introduction to take cell death into account, by adding a Poisson process of cell death which competes with the process of malignant transformation, as suggested by Yakovlev and Polig [29]. This model has been successfully applied to data from radiation induced and chemically induced lesions [30–32]. With the cell death component, it becomes less clear that the hazard function should increase monotonically in the case of spontaneous carcinogenesis. In fact, in the simplified case of constant rates ν_1 of initiation and ν_2 of cell death, and arbitrary cumulative distribution function $F(t)$ for time to transformation of intermediate lesions, the hazard function for time to tumor has the form

$$\lambda(t) = \nu_1 \exp(-\nu_2 t) F(t). \tag{15}$$

We note that according to this model the clock for cell death in this model starts at birth. If the constant $\nu_2 > 0$ in (15), then $\lambda(t)$ must decrease exponentially since $F(t)$ approaches one as t approaches infinity. We conjecture that in the present context the cell death component is very small, so that it does not dominate $\lambda(t)$ until after age 85. The higher hazard rate for male colorectal cancer, as well as the continued increase in hazard through age 105, may be attributed to a smaller rate of cell death. Another possibility is that the cell death should not be measured from birth, but from formation of the initiated cell (as in another variation of the model suggested in [29]).

We have noted in the Results section that proportionality of hazard appears to fail after age 90 for both breast and colorectal cancer in women. This result may be due to sampling variability, or additional bias unique to women at these high ages. We note that there are only 116 female breast cancer cases and 77 female colorectal cancer cases after age 90. They are distributed over a fifteen year period, for an average of 7.7 breast cancer and 5.1 colorectal cancer cases per year in this range. In addition, data linkage is more difficult for women, who are more likely to have changed names than men. An additional indication that the lack of proportionality for women may be spurious is that we do not see this apparent lack of proportionality in men.

REFERENCES

1. A.Y. Yakovlev, A.D. Tsodikov, K. Boucher and R. Kerber, The shape of the hazard function for breast cancer: Curability of the disease revisited, *Cancer* **85**, 1789–1798, (1999).
2. O.O. Aalen, Modeling heterogeneity in survival analysis by the compound Poisson distribution, *Annals of Applied Probability* **2**, 951–972, (1992).
3. D.F. Easton, The inherited component of cancer, *British Medical Bulletin* **50**, 527–535, (1994).
4. J.P. Struewing, P. Hartge, S. Wacholder, S.M. Baker, M. Berlin, M. McAdams, M.M. Timmerman, L.C. Brody and M.A. Tucker, The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews, *N. Engl. J. Med.* **336**, 1401–1408, (1997).
5. D. Ford and D.F. Easton, The genetics of breast and ovarian cancer, *Br. J. Cancer* **72**, 805–812, (1995).
6. B. Newman, H. Mu, L.M. Butler, R.C. Millikan, P.G. Moorman and M.C. King, Frequency of breast cancer attributable to BRCA1 in a population-based series of American women, *JAMA* **279**, 915–921, (1998).
7. T.I. Anderson, Genetic heterogeneity in breast cancer susceptibility, *Acta Oncol.* **35**, 407–410, (1996).
8. P. Vehmanen, L.S. Friedman, H. Eerola, L. Sarantaus, S. Pyrhönen, B.A.J. Ponder and T. Muhonen *et al.*, A low proportion of BRCA2 mutations in Finnish breast cancer families, *Am. J. Human Genet.* **60**, 1050–1058, (1997).
9. R.S. Houlston, A. Collins, J. Slack and N.E. Morton, Dominant genes for colorectal cancer are not rare, *Ann. Human Genet.* **56**, 99–103, (1992).

10. S.J. Laken, G.M. Petersen, S.B. Gruber, C. Oddoux, H. Ostrer and G.M. Giardiello *et al.*, Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC, *Nat. Gent.* **17**, 79–83, (1997).
11. S.H. Moolgavkar and D.J. Venzon, Two-event models for carcinogenesis: Incidence curves for childhood and adult tumors, *Math. Biosci.* **47**, 55–77, (1979).
12. S.H. Moolgavkar and A. Knudson, Mutation and cancer: A model for human carcinogenesis, *J. Natl. Cancer Institute* **66**, 1037–1052, (1981).
13. S.H. Moolgavkar and E.G. Luebeck, Two-event model for carcinogenesis: Biological, mathematical and statistical considerations, *Risk Anal.* **10**, 323–341, (1990).
14. W.F. Heidenreich, E.G. Luebeck and S.H. Moolgavkar, Some properties of the hazard function of the two-mutation clonal expansion model, *Risk Analysis* **17**, 391–399, (1997).
15. A. Kopp-Schneider, C.J. Portier and C.D. Sherman, The exact formula for tumor incidence in the two-stage model, *Risk Analysis* **14**, 1079–1080, (1994).
16. Q. Zheng, On the exact hazard and survival functions of the MVK stochastic carcinogenesis model, *Risk Anal.* **14**, 1081–1084, (1994).
17. L.G. Hanin and A.Y. Yakovlev, A nonidentifiability aspect of the two-stage model of carcinogenesis, *Risk Anal.* **16**, 711–715, (1996).
18. A.Yu. Yakovlev and A.D. Tsodikov, *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific, Singapore, (1996).
19. J.D. Kalbfleisch and R.L. Prentice, *The Statistical Analysis of Failure Time Data*, Wiley, New York, (1980).
20. T.R. Fleming and D.P. Harrington, *Counting Processes and Survival Analysis*, Wiley, New York, (1991).
21. N. Keiding, Independent delayed entry, In *Survival Analysis: The State of the Art*, (Edited by J.P. Klein and P.K. Goel), pp. 309–326, Kluwer, Boston, MA, (1992).
22. M. Woodroofe, Estimating a distribution function with truncated data, *Ann. Statist.* **13**, 163–177, (1985).
23. E.L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481, (1958).
24. N. Keiding and R.D. Gill, Random truncation models and Markov processes, *Ann. Statist.* **18**, 582–602, (1990).
25. S. Johansen, The product limit as a maximum likelihood estimator, *Scand. J. Statist.* **5**, 195–199, (1978).
26. M.-C. Wang, N.P. Jewell and W.-Y. Tsai, Asymptotic properties of the product limit estimate under random truncation, *Ann. Statist.* **14**, 1597–1605, (1986).
27. J. Etezadi-Amoli and A. Ciampi, Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function, *Biometrics* **43**, 181–192, (1987).
28. D.M. Himmelblau, *Applied Nonlinear Programming*, McGraw-Hill, Austin, (1972).
29. A. Yakovlev and E. Polig, A diversity of responses displayed by a stochastic model of radiation carcinogenesis allowing for cell death, *Math. Biosci.* **132**, 1–33, (1966).
30. A. Yakovlev, W. Müller, L. Pavlova and E. Polig, Do cells repair precancerous lesions induced by radiation?, *Math. Biosci.* **142**, 107–117, (1997).
31. K.M. Boucher and A.Y. Yakovlev, Estimating the probability of initiated cell death before tumor induction, *Proc. Natl. Acad. Sci. USA* **94**, 12776–12779, (1997).
32. K. Boucher, L.V. Pavlova and A.Y. Yakovlev, A model of multiple tumorigenesis allowing for cell death quantitative insight into biological effects of urethane, *Math. Biosci.* **150**, 63–82, (1998).

Journal of Epidemiology and Biostatistics (2001) Vol. 6, No. 5, 377–385

Measures of familial aggregation as predictors of breast-cancer risk

KM BOUCHER AND RA KERBER
Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah, Salt Lake City, UT, USA

Background Several measures of familial disease aggregation have been proposed, but only a few of these are designed to be implemented at the individual level. We evaluate two of them in the context of breast-cancer incidence.

Methods A population-based cohort consisting of 114 429 women born between 1874 and 1931 and at risk for breast cancer after 1965 was identified by linking the Utah Population Data Base and the Utah Cancer Registry. Two competing methods were used to obtain predictors of familial aggregation of risk: the number of first-degree relatives with breast cancer (N1ST) and the familial standardised incidence ratio (FSIR), which weights the disease status of relatives based on their degree of relatedness with the proband. Relative risks were estimated using Mantel–Haenszel, Poisson regression and spline regression methods. The age-dependent hazard function was also estimated.

Results Compared to a baseline category containing 91.5% of the subjects, the 0.7% of subjects identified as high risk using the FSIR criterion had a relative risk of about 2.8, while those identified as high risk using the N1ST criterion had a relative risk of 2.0. Moderate-risk subjects had a relative risk of about 1.75 using either criterion. FSIR was a significant predictor of risk even for those with no affected first-degree relatives. No decline in the baseline risk was observed at advanced ages.

Conclusions FSIR appears to be a better predictor of breast-cancer risk than N1ST, particularly for high-risk subjects.

Keywords familial risk, hazard function, truncation, survival analysis, breast cancer.

Introduction

Heterogeneity in a population may lead to population estimates of the hazard that do not reflect individual risk. For example, if we let $\lambda(t)$ denote the hazard function, and p the probability of immunity to a particular disease, it follows from the formula:

$$p = \lim_{t\to\infty} \exp\left\{-\int_0^t \lambda(u)du\right\} \qquad (1)$$

that there are individuals who are 'immune' in the population exactly when the hazard function has finite integral. In particular, $\lim_{t\to\infty}\lambda(t) = 0$, provided the limit exists. More generally, a large degree of heterogeneity in disease susceptibility may lead to a population hazard-function with one or more well-defined maxima. The maxima may correspond to discrete subpopulations with different genetic predisposition to disease. A maximum may also result from a continuous frailty, as the surviving population at higher ages may be over-represented by individuals with lower risk[1].

There is evidence of heterogeneity for most cancers. According to Easton[2], 'All cancer types exhibit familial clustering, suggestive of a significant inherited component'. He goes on to conclude that, as of 1994, known cancer genes accounted for 0.5–1% of all cancer cases, and that this figure would increase as more cancer genes are discovered. The breast-cancer genes *BRCA1* and *BRCA2* both contribute to an increased risk of breast cancer. *BRCA1* has an estimated allele frequency of 0.0002–0.001 [95% confidence interval (CI)][3], and accounts for about 3% of diagnosed breast cancer[4]. The allele frequency of mutations in *BRCA2* is estimated at 0.00022[5]. Vehmanen *et al.*[6] found that only 21% of breast-cancer families were accounted for by mutations of *BRCA1* and *BRCA2*, providing indirect evidence for the existence of other, undiscovered breast-cancer genes.

In our previous paper[7], linked population-based data from the Utah Cancer Registry (UCR) and the Utah Population Data Base (UPDB) was used to estimate the population-level hazard function for breast and colorectal cancer, stratified by birth cohort. We found that the hazard functions for both breast and colorectal cancer appeared to be monotone increasing functions for both genders and all birth cohorts. This contrasts with the model-based estimates of Moolgavkar *et al.*[8], who

Correspondence to: KM Boucher, Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112, USA.

Received 18 September 2000
Revised 9 May 2001
Accepted 26 June 2001

found the hazard function to decrease sharply starting sometime past the age of 70. The lack of clear multiple modes in the hazard function highlighted the fact that more delicate methods would be needed to account for the known heterogeneity of risk.

A number of measures of familial disease aggregation have been used or proposed, but only a few of these are designed to be implemented at the individual level. The most common epidemiological measure of familial risk is an indicator of whether one or more first-degree relatives have been diagnosed with cancer, or some other disease. Khoury and Flanders[9] have noted that measures of this sort are prone to bias under a variety of conditions. Nonetheless, it is a widely used and easily understood measure of familial risk that can easily be ascertained in a clinical setting. A second category of family-history measures, suggested by Kerber[10], is derived from the complete risk experience of all observable biological relatives, adjusted for the age, sex, number and degree of the relatives. The total familial risk is summarised as a familial standardised incidence ratio (FSIR), or a familial rate (FR). FSIR and FR are less prone to bias and substantially more sensitive than a crude indicator variable, but require fairly detailed family history data, which may rarely be available in a clinical setting.

In this paper, age-specific estimates of the relative risk (RR) for breast cancer are obtained using three different methods, with the number of affected first-degree relative (N1ST) and FSIR as predictors of risk. We show that FSIR performs better in identifying subjects at high risk. As a by-product of the analysis we find that the risk for breast cancer appears to be increasing as a function of age, even at advanced ages.

Data

The data for this study were obtained by linking records from the UPDB with the UCR. The UPDB consists of the genealogical records of more than 1 000 000 individuals who were born, died, or married in Utah, or en route to Utah, during the nineteenth and twentieth centuries. The available follow-up information comes either from Utah death certificates, which have been linked to the UPDB genealogical data every year from 1933 through the beginning of 1997, or from linkage of the HCFA beneficiary data to the UPDB. The study population consisted of 122 208 women recorded in the Utah Population Database, who were born during 1874–1931 and for whom follow-up information was available that places them in Utah during the years of operation of the Utah Cancer Registry (1966–present). Subjects with purported follow-up past age 105 were excluded from the data. Potential subjects who had no relatives who were also in the risk set, and therefore for whom no measures of familial aggregation could be computed, were removed from the data. Excluding these two groups removed an additional 7779 women, leaving a study population of 114 429 women. There are 5092 cases of female breast cancer in the data. Only female breast cancer was analysed. Additional details on the data are given in Boucher and Kerber[7].

Methods

Number of first-degree relatives

The simplest and most easily understandable measure of familial aggregation is the number of affected N1ST. Of the 114 429 women in the data set, 9765, or approximately 8.5%, had at least one N1ST with breast cancer and 795 women, or 0.69%, had two or more affected N1ST. Having more than two N1ST with breast cancer was extremely rare: 56 women had three, and 10 women had the maximum of four. For this reason, subjects with two or more affected relatives were combined for data analysis. Table 1 gives the distribution of number of cases and person-years of risk, stratified by N1ST, age, and birth-year. Mantel–Haenszel RR and 95% CI are also presented.

Familial standardised incidence ratio

The second measure of familial aggregation we used was a modification of the FSIR[10]. The FSIR is derived from the complete risk experience of all observable biological relatives, adjusted for age, sex, number and shared inheritance with the subject. Formally, FSIR is defined in terms of the kinship coefficient[11] $c(i,j)$ between individuals i and j, which gives the probability that two individuals share randomly selected genes at a given locus by common descent. The kinship coefficient is defined by

$$c(i,j) = (1/2)\sum_{p=1}^{P_{i,j}} 2^{-l(p)}$$

where $P_{i,j}$ is the total number of distinct shortest paths through a common ancestor between individuals i and j, and $l(p)$ is the length in number of reproductive events of the path p.

A simple case is the one in which the subjects with indices i and j are full siblings. In this case, there are two paths of length two between the subjects, one through the father and one through the mother, and $c(i,j) = (1/2)(2^{-2} + 2^{-2}) = 0.25$. For another common example, consider the case in which the individuals labeled i and j are first cousins. In the most typical case, there is exactly one pair of shared grandparents, with the other pairs of grandparents distinct and unrelated to each other or the shared pair. In this case, there are two paths

Table 1 Distribution of cases (C) and person-years (PY) stratified by age and N1ST

Age	N1ST										
	0	1					2+				
	C/PY	C/PY	RR_c	RR_a	95% CI		C/PY	RR_c	RR_a	95% CI	
					Lower	Upper				Lower	Upper
35–39	13/ 31509	5/ 2976	4.07	3.73	1.30	10.68	1/ 292	8.30	6.40	0.66	62.06
40–44	55/ 89836	10/ 8663	1.89	1.75	0.88	3.50	0/ 803	0.00	0.00	-	-
45–49	158/ 154308	24/ 15130	1.55	1.48	0.95	2.29	5/ 1401	3.49	3.56	1.48	8.60
50–54	192/ 218367	50/ 21591	2.63	2.39	1.74	3.30	1/ 2041	0.56	0.65	0.10	4.02
55–59	384/ 277099	60/ 27254	1.59	1.61	1.23	2.11	7/ 2585	1.95	1.83	0.85	3.97
60–64	552/ 329333	89/ 31454	1.69	1.75	1.41	2.19	7/ 2944	1.42	1.42	0.68	2.99
65–69	730/ 348857	131/ 32422	1.93	1.91	1.58	2.30	14/ 2992	2.24	2.34	1.40	3.91
70–74	744/ 317980	103/ 28408	1.55	1.51	1.23	1.86	14/ 2584	2.32	2.05	1.18	3.54
75–79	642/ 264036	105/ 22007	1.96	1.86	1.51	2.29	12/ 1949	2.53	2.43	1.37	4.29
80–84	470/ 190439	63/ 14468	1.76	1.64	1.25	2.13	10/ 1201	3.37	3.09	1.65	5.79
85–89	267/ 110097	25/ 7577	1.36	1.19	0.78	1.82	1/ 605	0.68	0.11	0.001	11.13
90–94	99/ 45774	13/ 2874	2.09	1.89	1.05	3.43	0/ 211	0.00	0.00	-	-
95+	31/ 12083	5/ 644	3.03	3.20	1.26	8.11	0/ 30	0.00	0.00	-	-
All	4337/2389718	683/215468	1.75	1.75	1.61	1.89	72/19638	2.02	2.01	1.59	2.54

Crude RR (RR_c), Mantel–Haenszel birth-year adjusted RR (RR_a) and 95% CI are also presented. The reference group for each age category is the lowest level of N1ST (N1ST = 0). The adjusted RR values in the final row are adjusted for age and birth-year.

of length four between individuals i and j and $c(i,j) = (1/2)(2^{-4}+2^{-4}) = 1/16$. The relevant reproductive events in this case are the ones corresponding to subjects i and j, and the reproductive events for the parent of subject i and the parent of subject j that are siblings. Finally, we suppose that we have a stratified population; the number of strata is K, the population incidence in the kth stratum is given by λ_k, and let t_{jk} be the time that the jth person spent in the kth stratum of risk. Let $I_j = 1$ if the jth member has the disease and 0 otherwise. The FSIR is then defined, for the ith individual, by:

$$FSIR_i = \frac{\sum_{j \neq i} I_j c(i,j)}{\sum_{k=1}^{K} \sum_{j \neq i} t_{jk} \lambda_k c(i,j)} \qquad (2)$$

The sum for the index j runs over all related individuals in the pedigree (excluding subject i).

In deriving a measure of variance VAR_i for $FSIR_i$, it was assumed that the denominator of the above expression is fixed, and that for each fixed path length the number of observed cases follows a Poisson distribution with mean equal to the expected number of cases in the stratum. For this study, risk was stratified by age and gender.

A difficulty with using the 'raw' FSIR scores is that the amount of information from which it is constructed for a particular individual is highly variable. A low FSIR score could be an indicator of low risk, or simply reflect a small family size. We therefore chose to adjust the scores using an empirical Bayes procedure before incorporating them into a regression analysis. As the raw FSIR scores are highly skewed, we applied the empirical Bayes procedure to $\log(1+\log(1+FSIR))$. The basic assumption of the empirical Bayes adjustment is that the 'true' values μ_i of $\log(1+\log(1+FSIR_i))$ are normally distributed. The mean and variance of μ are estimated empirically and iteratively from the data. The procedure we use is similar to the one suggested by Greenland and Robins[12].

More specifically, we suppose that after iteration $n-1$ we have current estimates $\mu_{i,n-1}$ and $\sigma_{i,n-1}$ for the true value of $\log(1+\log(1+FSIR))$ for the ith individual, as well as an overall mean μ_{n-1} and variance σ_{n-1} for the μ_i. We then computed new estimates using the formula

$$\mu_{i,n} = \mu_{n-1} + \left(\frac{\sigma^2_{n-1}}{\sigma^2_{n-1} + \sigma^2_{i,n-1}}\right)(Y_i - \mu_{n-1}) \qquad (3)$$

where $Y_i = \log(1+\log(1+\mathrm{FSIR}_i))$, with variance estimated by:

$$\sigma^2_{i,n} = \frac{VAR_i}{(\exp(\mu_{i,n-1})\exp(\exp(\mu_{i,n-1})-1))^2} \qquad (4)$$

given by the delta method. We then computed the sample mean and variance of $\mu_{i,n}$, over all the subjects to get μ_n and σ^2_n. The iteration was repeated until the values stabilised.

The distribution of $\log(1+\log(1+FSIR))$, before and after transformation, is displayed in Figure 1. Note that the 'raw' distribution is bimodal, with a mode at zero that disappears after transformation.

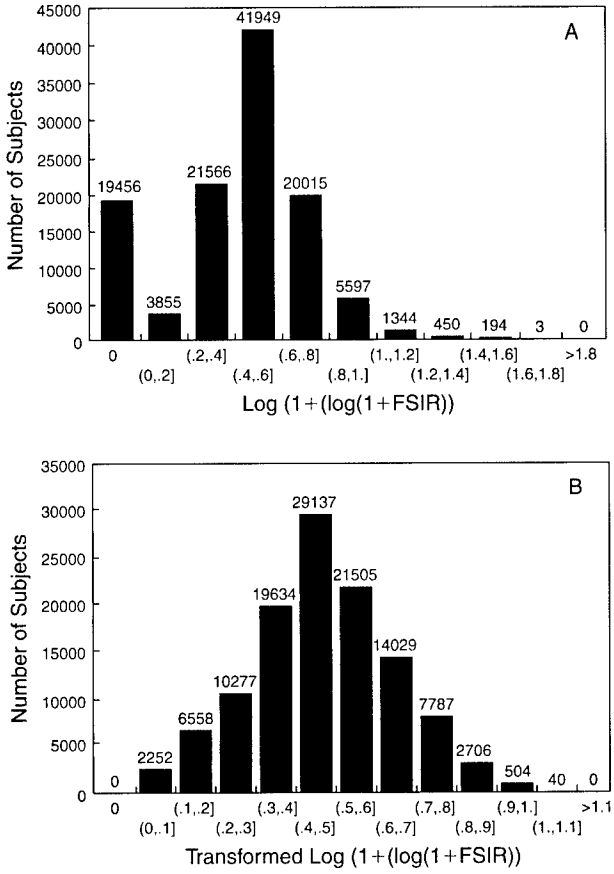


Fig. 1 *The distribution of log(1+log(1+FSIR)) before (**A**) and after (**B**) empirical Bayes adjustment.*

The inverse of the transformation $\log(1+\log(1+FSIR))$ was then used to adjust FSIR. We then divided the adjusted FSIR into 'Low', 'Medium' and 'High' risk categories, containing the same fraction of the data as the corresponding categories of N1ST. Table 2 gives the distribution of cases and person-years, stratified by adjusted FSIR, age and birth-year. Mantel–Haenszel RR and 95% CI are also presented.

Regression methods

Two regression methods were used to analyse the data. The first was a standard Poisson regression method, using the grouped data presented in Tables 1 and 2. The data was further stratified into 5-year birth-year intervals. Let $d_{i,j,k}$ and $n_{i,j,k}$ be the number of cases and person-years of risk for stratum (i,j,k), where i indexes age, j indexes birth-year, and k indexes familial risk (either N1ST or adjusted FSIR). Let BY_j be the midpoint of the jth birth-year category. The final Poisson regression models took the form:

$$\log(d_{i,j,k}) = \log(n_{i,j,k}) + a_i + \beta\log(BY_j) + \gamma_k \qquad (5)$$

A model was also considered with interaction terms $\eta_{i,k}$ between age and familial risk. The interaction terms were found to be insignificant both with N1ST ($\chi^2 = 31.14$ on 22 degrees of freedom, $p = 0.093$) and with adjusted FSIR ($\chi^2 = 20.68$ on 23 degrees of freedom, $p = 0.60$) as measures of familial risk. We therefore chose for simplicity to leave the interaction term out of our final model.

Since individual level data were available for this study, survival analysis methods were appropriate, and potentially more efficient than methods for aggregate data. In addition, it was possible to handle predictors in a truly continuous fashion, and obtain a smooth estimate of the age-dependent risk. We wished for the analysis to be essentially nonparametric, we therefore modelled the hazard via a quadratic spline[13]. A quadratic spline takes the form of a polynomial of second degree between successive break points τ_j, called 'knots'. A quadratic spline with m knots specifies the hazard to be of the form:

$$h_m(t) = \sum_{i=0}^{2}\gamma_{0i}t^i + \sum_{j=1}^{m}\gamma_{j2}(t-\tau_j)_+^2 \qquad (6)$$

where $(\chi)_+ = \max(\chi,0)$. The spline is a continuous function with continuous first derivative.

We fit splines with knots that were equally spaced in the interior of the interval $[T_{min},T_{max}]$, where T_{min} is the minimum age of any subject in 1965, and T_{max} the maximum follow-up (failure or censoring) time. Thus, with m knots the number of parameters was $m+3$. As the number of knots increases, the fit becomes increasingly nonparametric. Since the number of knots was not specified in advance, the analysis was nonparametric in character. The data described are subject to random truncation: cases that occurred during or before 1965 are not recorded in the dataset. Subjects were between the ages of 34–86 at the time of truncation. In the more familiar clinical setting in which survival analysis is used, time is set to zero for all patients at the time of diagnosis or treatment, so truncation is not an issue. The truncation is taken into account in Poisson regression and in the calculation of FSIR scores, by appropriately adjusting the number of subjects in the risk-set. Thus, analysis of the data must take into account not only the effects of right censoring, but also the effects of left truncation due to delayed entry into the risk-set. The truncation was handled by using a conditional likelihood; conditioning on the event that the age at breast cancer was greater than the age at which as subject entered the risk-set. The spline regression models all contained the logarithm of birth-year as a continuous covariate, in addition to a measure of familial risk.

More specifically, let X be the observed failure or censoring age, let Y be the age at truncation and let δ be a censoring indicator ($\delta=1$ if the failure is observed and $\delta=0$ otherwise). Let $\vec{s}$ be a vector of covariates and $\vec{z}$ a vector of parameters. Let $\lambda(x,\vec{s};\vec{z})$ denote the hazard associated with the failure time-distribution $F(x,\vec{s};\vec{z})$ and $\Lambda(x,\vec{s};\vec{z})$ the cumulative hazard for $F(x,\vec{s};\vec{z})$. Let i index the subjects. The likelihood, conditional on $X>Y$, becomes:

$$\log(CL) = \sum_i[\delta_i \log(\lambda(x_i,\vec{s_i};\vec{z_i})) - (\Lambda(x_i,\vec{s_i};\vec{z_i}) - \Lambda(y_i,\vec{s_i};\vec{z_i}))] \qquad (7)$$

The proportional hazards model $\lambda(x,\vec{s};\vec{z}) = \exp(\vec{s}^t\vec{z})\lambda_0(x)$ was used to model covariate effects (the logarithm of birth year and family history).

The spline estimates were computed by maximizing the logarithm of the conditional likelihood using the algorithm of Powell[14]. We started with one knot and increased the number of knots until the fit was not improved, using the log-likelihood as a guide. This occurred when there were three knots. We thus used three knots for all the results in the following section.

Results

Estimates of RR

Tables 3 and 4 present estimates of the RR and 95% CI for 'Moderate' and 'High' risk categories of N1ST and FSIR, compared with the 'Low' risk category, using each of the three methods presented in the previous section. The Mantel–Haenszel, Poisson, and spline-based estimates are in remarkably good agreement. We see that the 'Moderate' risk categories of both N1ST and FSIR confer an approximate 1.75-fold increased risk. The 'High' risk category of N1ST confers an approximate 2.0-fold increased risk, while the 'High' risk category of FSIR confers an approximate 2.8-fold RR. FSIR appeared to identify subjects with higher than the 'Moderate' level of risk. The 'High' and 'Moderate' risk categories of N1ST, on the other hand, have almost the same estimated risk. As judged by the length of the CI, there

Table 2 Distribution of cases (C) and person-years (PY) stratified by age and FSIR

Age	Adjusted FSIR										
	0.01–1.8	1.8–3.1					3.1–6.0				
	C/PY	C/PY	RR_c	RR_a	95% CI		C/PY	RR_c	RR_a	95% CI	
					Lower	Upper				Lower	Upper
35–39	13/ 31851	5/ 2643	4.64	4.29	1.50	12.28	1/ 283	8.66	6.63	0.68	64.69
40–44	57/ 90953	8/ 7650	1.67	1.58	0.74	3.36	0/ 699	0.00	0.00	-	-
45–49	157/ 156426	27/ 13240	2.03	1.80	1.18	2.76	3/ 1173	2.55	3.09	1.09	8.78
50–54	194/ 221515	44/ 18839	2.67	2.26	1.60	3.19	5/ 1645	3.47	3.97	1.72	9.14
55–59	387/ 280996	57/ 23846	1.74	1.65	1.25	2.20	7/ 2096	2.42	2.46	1.17	5.16
60–64	566/ 333467	71/ 27776	1.51	1.55	1.22	1.98	11/ 2488	2.60	3.02	1.73	5.27
65–69	744/ 352061	116/ 29621	1.85	1.82	1.50	2.22	15/ 2589	2.74	2.76	1.66	4.61
70–74	742/ 319737	106/ 26876	1.70	1.72	1.41	2.11	13/ 2359	2.37	2.26	1.28	3.99
75–79	644/ 263876	97/ 22147	1.79	1.73	1.39	2.15	18/ 2969	3.75	3.84	2.41	6.14
80–84	468/ 189091	64/ 15682	1.65	1.65	1.27	2.15	11/ 1335	3.33	3.55	1.98	6.35
85–89	257/ 108566	32/ 8984	1.50	1.46	1.01	2.12	4/ 729	2.32	1.42	0.41	4.90
90–94	92/ 44798	19/ 3804	2.43	2.58	1.59	4.18	1/ 257	1.89	1.49	0.16	13.83
95+	33/ 11815	3/ 886	1.21	1.28	0.39	4.18	0/ 56	0.00	0.00	-	-
All	4354/2405152	649/201994	1.77	1.74	1.60	1.89	89/17678	2.78	2.89	2.35	3.56

Crude RR (RR_c), Mantel–Haenszel birth-year adjusted RR (RR_a) and 95% CI are also presented. The reference group for each age category is the lowest level of FSIR. The adjusted RR in the final row are adjusted for age and birth-year.

Table 3 RR, 95% CI, and χ^2 values for N1ST, computed using the Mantel–Haenszel, Poisson regression and a spline regression method

Category	Model					
	Mantel–Haenszel		Poisson		Spline	
	Estimate	MH χ^2	Estimate	Wald χ^2	Estimate	LR χ^2
N1ST = 0	1	-	1	-	1	-
N1ST = 1	1.75 (1.61–1.89)	13.80	1.76 (1.62–1.90)	186.91	1.75 (1.61–1.90)	-
N1ST = 2+	2.01 (1.59–2.54)	6.00	2.04 (1.61–2.57)	35.82	2.03 (1.58–2.56)	186.92[a]

See the text for details. All the methods adjust for birth-year and age and have the lowest level as reference category.
[a]Two df likelihood ratio χ^2 for inclusion of both (N1ST = 1) and (N1ST = 2).

Table 4 RR, 95% CI, and χ^2 values for FSIR, computed using the Mantel–Haenszel, Poisson regression and a spline regression method

Category	Model					
	Mantel–Haenszel		Poisson		Spline	
	Estimate	MH χ^2	Estimate	Wald χ^2	Estimate	LR χ^2
Low FSIR	1	-	1	-	1	-
Moderate FSIR	1.74 (1.60–1.89)	13.92	1.67 (1.65–1.93)	188.77	1.78 (1.65–1.93)	-
High FSIR	2.89 (2.35–3.56)	10.05	2.83 (1.17–3.46)	93.99	2.83 (2.27–3.46)	227.02[a]

See the text for details. All the methods adjust for birth-year and age and have the lowest level as reference category.
[a]Two df likelihood ratio χ^2, or inclusion of both 'Moderate FSIR' and 'High FSIR'.

appears to be little gain in efficiency for estimation of the RR from the regression method using splines.

More than 90% of subjects have no affected first-degree relatives, and thus must be lumped together in the 'Low' risk category. To explore the affect of this categorisation further, we fit a model with FSIR treated as a continuous predictor, as well as log(birth-year), to the subset consisting only of those subjects with N1ST = 0. For comparison, we fit a similar model to the entire dataset. FSIR was a highly significant predictor of risk, even when restricted to the subjects with no affected first-degree relative, with significance $p < 0.00001$ as measured by the likelihood ratio test (with the logarithm of birth-year also in the reference model). The parameter estimate for FSIR from the restricted data was 1.11 (95% CI 0.91–1.31). This model corresponds to an approximate three-fold range of risk in the subjects with N1ST = 0. By comparison, the parameter estimate for the entire data set was 1.45 (95% CI 1.29–1.62), which corresponds to an approximate 5-fold range of risk over the entire dataset. Thus, there appears to be a significant range of familial risk among the subjects with no affected first-degree relatives. The estimated range of risk is higher using FSIR as a continuous variable.

Estimation of the age-dependent risk

Estimates of the age-dependent risk for breast cancer for the 1901–05 birth cohort are presented in Figure 2. The estimates for the 'Low' level of risk using N1st and FSIR are essentially identical. The estimates from Poission regression and the spline model are in good agreement. Figure 3 shows the effect of birth-year on the 'Low' category of risk, using only the spline models containing FSIR. The crude estimates of risk are also presented in Figure 3 for comparison. Adding one to the birth-year corresponded to an approximate 2.7% increase in risk (95% CI 2.4–3.1%), with FSIR in a Poisson regression model. Using the spline model with FSIR the estimated increase in risk per year is 3.5% (95% CI 3.2–3.9%). Some of this difference may be due to the slightly different way in which birth-year was incorporated into the models.

The crude estimate shows a decrease in risk near ages 50–55 that all but disappears after adjusting for birth-year. There is also a less pronounced decrease in the crude risk near age 90 that also disappears after adjustment for birth-year. There is little evidence for a decrease in risk up to the age of 95 or 100. More recent birth cohorts appear to be at significantly increased risk compared to earlier birth cohorts.

Discussion

We have applied two methods of measuring familial aggregation at the individual level to breast cancer data. As might be expected, both prove to be highly significant predictors of individual risk of breast cancer. FSIR

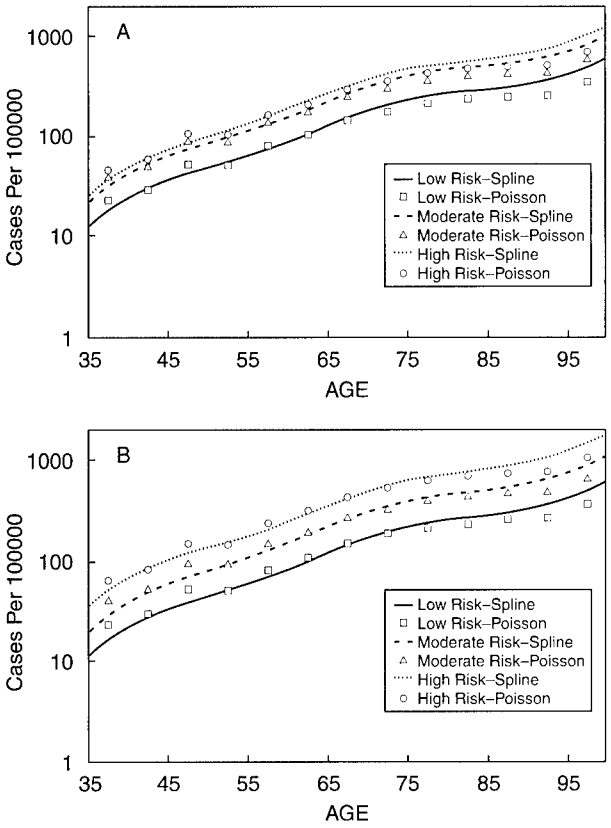


*Fig. 2 The age-dependent risk for breast cancer for the 1901–05 birth cohort, estimated from Poisson regression and a spline regression model. Estimates are presented using N1ST (**A**) and FSIR (**B**) as predictors of familial risk.*

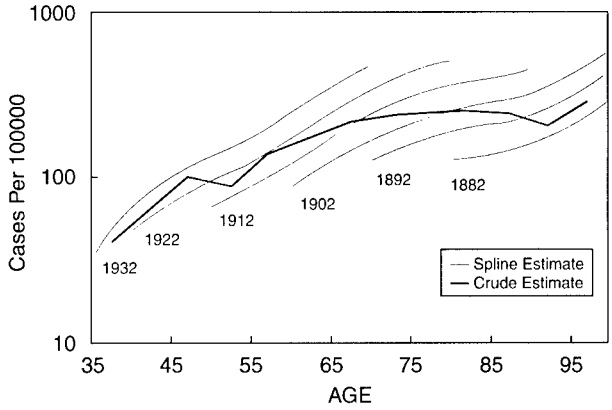


Fig. 3 The age-dependent risk for the lowest level of FSIR, estimated for several birth cohorts using a spline regression model. The crude estimate from the tabulated data, which is not adjusted for birth-year, is also presented for comparison.

appeared to perform better, however, at identifying small fractions of subjects at the highest risk. This also may not be surprising, since FSIR is adjusted for the expected number of cases in the relatives of the subject.

In addition, by performing a restricted analysis using FSIR as a predictor in the subset of subjects with no affected first-degree relatives, we see that N1ST does not appear to contain enough information to separate out individuals at increased risk. Thus, a measure that incorporates the observed risk of a wider class of relatives seems warranted.

Methods for aggregate data, such as the Poisson regression method, perform remarkably well compared to the spline regression we used, as long as the data is categorised similarly in the models. Any advantage for the spline model appears to be in the ability to treat variables in a truly continuous way. These advantages allow for more complicated modeling strategies, not exploited in the present paper, such as accelerated failure time models.

As an interesting by-product of the analysis, we find that, with the exception of a possible slight decrease in risk between the ages of 50 and 60, the risk for breast cancer appears to be non-decreasing through age 95. Thus, we find no evidence of an 'immune fraction' for breast cancer. First, the lowest level of N1ST appears to have a high degree of variability of familial risk. Secondly, the age-dependent risk is the lowest category of N1ST or FSIR and shows no evidence of a peak in risk followed by a drop, which would be the case if susceptible individuals were contracting breast cancer and being removed from the population, thus increasing the immune fraction in the surviving population.

Other investigators have either estimated or simply assumed that the risk of breast cancer decreases past a certain age. As previously noted, Moolgavkar *et al.*[8], found the hazard function to decrease sharply starting sometime past the age of 70. By age 90, the risk has decreased to about 1/3 of the peak. Parmigiani *et al.*[15] fit breast-cancer incidence data from Easton *et al.*[16] to a three parameter gamma distribution. Implicit in this fitting procedure is the assumption that the risk to carriers of *BRCA1* and *BRCA2* decreases to zero with age. There is little evidence for this in the data used by Parmigiani *et al.*, as the highest age is 70. It may be important for further efforts at risk prediction to better understand the hazards to carriers of disease susceptibility genes, particularly at more advanced ages, where data are sparse.

Acknowledgements

This research was supported, in part, by NCI Cancer Center Support Grant 2P30 CA 42014, US Army Medical Research and Material Command Grant DAMD17-1-8256, and by NIH/NCI grant R29 CA69421. The Huntsman Cancer Institute provided partial support for the Utah Population Data Base. The Utah Cancer Registry was supported by NCI grant NO1

PC 67000. In addition, the authors would like to thank Professor Andrei Y. Yakovlev for many helpful discussions.

References

1 Aalen OO. Modeling heterogeneity in survival analysis by the compound Poisson distribution. *Ann Appl Probabil* 1992;2:951–72.

2 Easton DF. The inherited component of cancer. *Br Med Bull* 1994;50;527–35.

3 Ford D, Easton DF. The genetics of breast and ovarian cancer. *Br J Cancer* 1995;72:805–12.

4 Newman B, Mu H, Butler LM *et al.* Frequency of breast cancer attributable to *BRCA1* in a population-based series of American women *J Am Med Assoc* 1998;279:915–21.

5 Anderson TI. Genetic heterogeneity in breast cancer susceptibility. *Acta Oncol* 1996;35:407-410.

6 Vehmanen P, Friedman LS, Eerola H *et al.* A low proportion of *BRCA2* mutations in Finnish breast cancer families. *Am J Hum Genet* 1997;60:1050–8.

7 Boucher KM, Kerber RA. The shape of the hazard function for cancer incidence. *Math Comput Model* In press.

8 Moolgavkar SH, Stevens RG , Lee JAH. Effect of age on incidence of breast cancer in females. *J Natl Cancer Inst* 1979;62:493–501

9 Khoury M, Flanders WD. Bias in using family history as a risk factor in case-control studies of disease. *Epidemiology* 1995;6:511–19.

10 Kerber RA. Method for calculating risk associated with family history of a disease. *Genet Epidemiol* 1995;12:291–301.

11 Malecot G. *Les mathematiques de l'hereditie* Paris: Masson, 1948.

12 Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* 1991;2:244–51.

13 Etezadi-Amoli J, Ciampi A. Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function. *Biometrics* 1987;43:181–92.

14 Himmelblau DM. *Applied nonlinear programming* Austin: McGraw-Hill, 1972.

15 Parmigiani G, Berry DA, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes *BRCA1* and *BRCA2*. *Am J Hum Genet* 1998;62:145–58.

16 Easton DF, Ford D, Bishop DT, Breast Cancer Linkage Consortium. Breast and ovarian cancer incidence in *BRCA1*-mutation carriers. *Am J Hum Genet* 1995;56:265–71.